UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Ciências
Câmpus Bauru

PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MÁQUINAS DE BOLTZMANN EM PROFUNDIDADE PARA RECONHECIMENTO DE EVENTOS EM VÍDEOS

BAURU
2021

R688m

Roder, Mateus
Máquinas de Boltzmann em Profundidade para Reconhecimento de Eventos em Vídeos / Mateus Roder. -- Bauru, 2021
63 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru
Orientador: João Paulo Papa
Coorientador: André Luis Debiaso Rossi

1. Inteligência artificial. 2. Aprendizado em Profundidade. 3. Máquinas de Boltzmann. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Mateus Roder

Máquinas de Boltzmann em Profundidade para Reconhecimento de Eventos em Vídeos

Dissertação de mestrado para o curso de Pós-Graduação em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus Bauru.
Financiadora: FAPESP - Proc. 2019/07825-1

Banca Examinadora

Prof. Dr. João Paulo Papa
Orientador

Prof. Dr. Jurandy Gomes de Almeida Junior

Prof. Dr. Antonio Carlos Sementille

Bauru, 25 de fevereiro de 2021.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Bauru

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE MATEUS RODER, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 25 dias do mês de fevereiro do ano de 2021, às 14:00 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de MATEUS RODER, intitulada **MÁQUINAS DE BOLTZMANN EM PROFUNDIDADE PARA RECONHECIMENTO DE EVENTOS EM VÍDEOS**. A Comissão Examinadora foi constituida pelos seguintes membros: Prof. Dr. JOAO PAULO PAPA (Orientador(a) - Participação Virtual) do(a) IBILCE / UNESP/Bauru (SP) , Prof. Dr. JURANDY GOMES DE ALMEIDA JUNIOR (Participação Virtual) do(a) Instituto de Ciência e Tecnologia / Universidade Federal de São Paulo, Prof. Dr. ANTONIO CARLOS SEMENTILLE (Participação Virtual) do(a) Departamento de Computação / UNESP/Câmpus de Bauru . Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final:_ APROVADO _ _ _ . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

João Paulo Papa

Prof. Dr. JOAO PAULO PAPA

Agradecimentos

Agradeço à Deus pelas oportunidades e pessoas que colocou em minha vida, aos meus pais, à Amanda, minha namorada, e meus amigos, que sempre estiveram ao meu lado e torceram por mim.

À FAPESP, que aceitou financiar este trabalho, e à Unesp, instituição que faz parte da minha formação acadêmica e pessoal desde a graduação.

A todos os professores que tive contato neste período de muita agregação de conhecimento, e especialmente ao meu orientador, e amigo, professor João Paulo Papa, que me acolheu e auxiliou muito nesta caminhada. Também ao meu co-orientador, e amigo, professor André Luis Debiaso Rossi, que está presente em minha formação há anos.

Aos meus amigos do laboratório Recogna, que me acolheram e fizeram parte de inúmeros momentos, nestes dois anos de estudos. Especialmente, agradeço ao Gustavo Rosa, Leandro Passos, Luis Félix, Clayton Pereira, Douglas Rodrigues, e Claudio Santos, pelas parcerias e amizades construídas.

"Para criaturas pequenas como nós, a vastidão só é suportável por meio do amor."
(Carl Edward Sagan)

Resumo

Na última década, o crescimento exponencial dos dados apoiou o desenvolvimento de uma vasta gama de algoritmos baseados em aprendizado de máquina, além de possibilitar seus usos em aplicações cotidianas. Além disso, esta melhoria ou crescimento é parcialmente explicada pelo advento de técnicas de aprendizado em profundidade, ou seja, a composição de arquiteturas simples que geram modelos complexos e robustos. Embora técnicas de aprendizado em profundidade produzam resultados excelentes, elas também apresentam desvantagens em relação ao processo de aprendizagem, pois o treinamento de modelos complexos em grandes conjuntos de dados é computacionalmente custoso. Esse problema fica evidente quando se trata de análise e processamento de vídeos, como reconhecimento de ações ou eventos, uma vez que sequências de imagens (*frames*) são consideradas e produzem, geralmente, uma única saída. Outro problema relevante diz respeito à baixa quantidade de bancos de dados para determinadas tarefas, como a classificação de eventos de alto nível, fato que dificulta o desenvolvimento de algumas vertentes conceituais. Alguns trabalhos consideram a transferência de aprendizado ou a adaptação de domínio, ou seja, abordagens que mapeiam o conhecimento de um domínio para outro, a fim de aliviar a carga de treinamento, mas a maioria deles opera em blocos individuais ou pequenos blocos de *frames*. Portanto, neste trabalho é proposta uma nova abordagem para mapear o conhecimento entre domínios, do reconhecimento de ações até o reconhecimento/classificação de eventos utilizando modelos baseados em energia como função de mapeamento. Ademais, é proposta uma modificação no processamento dos vídeos para os modelos empregados, capaz de processar uma maior quantidade de *frames* simultaneamente, carregando informações espaciais e rastros temporais durante o processo de aprendizagem, o qual é denominado de processamento *Somatório*. Os resultados experimentais conduzidos em dois conjuntos de dados de vídeos públicos, o UCF-101 e o HMDB-51, retratam a eficácia da abordagem de adaptação de domínio e do processamento *Somatório* propostos, possibilitando uma redução do custo computacional em comparação aos modelos tradicionais baseados em energia, tais como Máquinas de Boltzmann Restritas, Redes de Crenças Profundas e Máquinas de Boltzmann Profundas.

Palavras-chave: Aprendizado em profundidade. Máquinas de Boltzmann Restritas. Classificação de eventos. Vídeos.

Abstract

In the past decade, the exponential growth of data has supported the development of a wide range of algorithms based on machine learning, enabling its uses in daily basis activities. Besides, such improvement is partially explained due to the advent of deep learning techniques, i.e., the composition of simple architectures that generate complex and robust models. Although both factors produce excellent results, they also have disadvantages concerning the learning process, since training complex models in large data sets are computationally expensive and time-consuming. This problem becomes evident when it comes to the video analysis and processing, as recognition of actions or events, since sequences of images (frames) are considered and usually generate a single output. Another relevant problem concerns the low number of high-level events classification databases, making it difficult to develop some conceptual aspects. Some studies consider transferring learning or a domain adapting, that is, approaches that map knowledge from one domain to another, to lighten the training load as most of them operate in individual blocks or small blocks of frames. Therefore, this work proposes a new approach to map knowledge between domains, from action recognition to event recognition/classification using energy-based models as a mapping function. Also, it is proposed a modification in the video processing for the employed models, capable of processing all frames simultaneously by carrying spatial and temporal information during the learning process, denoted as *Somatório* processing. The experimental results conducted over two public video data sets, the UCF-101 and the HMDB-51, portrait the effectiveness of the domain adaptation approach and the proposed *Somatório* models, reducing the computational load when compared to the standard energy-based models, such as Restricted Boltzmann Machines, Deep Belief Networks, and Deep Boltzmann Machines.

Key-words: Deep Learning. Restricted Boltzmann Machines. Event classification. Videos.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM	Aprendizado de Máquina
BN	*Bayesian Networks*
BRBM	*Bernoulli Restricted Boltzmann Machines*
CD	*Contrastive Divergence*
CNNs	*Convolutional Neural Networks*
CRF	*Conditional Random Fields*
DBM	*Deep Boltzmann Machines*
DBN	*Deep Belief Networks*
DL	*Deep Learning*
$\mathcal{D}_S$	*Source Domain*
$\mathcal{D}_T$	*Target Domain*
FC	*Fully-Connected*
HMM	*Hidden Markov Models*
MCMC	*Markov Chain Monte Carlo*
RBM	*Restricted Boltzmann Machines*
S-DBN	*Somatório Deep Belief Networks*
SGD	*Stochastic Gradient Descent*
S-RBM	*Somatório Restricted Boltzmann Machines*
TA	Transferência de Aprendizado

Sumário

1 Introdução

Frente aos grandes avanços tecnológicos, a sociedade encontra-se sob a nova era da informação, fortemente apoiada pela 4^{a} Revolução Industrial, caracterizada pela elevada conectividade entre sistemas ciber-físicos em ambientes produtivos industriais. Concomitantemente, criam-se e agregam-se diferentes tecnologias para a integração e ampliação do uso destas na sociedade, como sistemas de visão computacional baseados em inteligência artificial, frente à grande quantidade de dados gerada pelo aumento da conectividade entre os mais diferentes dispositivos (LI; HOU; WU, 2017).

No âmbito da visão computacional, busca-se a produção de representações de alto nível e intrínsecas do mundo real, de forma que essas características possibilitem a execução de tarefas como detecção e/ou classificação de objetos de maneira satisfatória por técnicas de aprendizado de máquina (BISHOP, 1995). Uma vez que o mundo real não é tão simples de ser parametrizado, tais tarefas tornam-se mais complexas quando existem variações ambientais de luminosidade, diferentes perspectivas e/ou planos de captura de imagens, variações na resolução, entre outros (LECUN; KAVUKCUOGLU; FARABET, 2010).

Tratando-se das abordagens tradicionais de aprendizado de máquina estas procuram resolver majoritariamente problemas de classificação a partir da extração de características de imagens, e posteriormente, utilizá-las para treinar um algoritmo de AM.

Entretanto, técnicas de aprendizado em profundidade (*Deep Learning* - DL) ganharam grande destaque e foco de estudo da comunidade científica (LECUN et al., 1998; HINTON; OSINDERO; TEH, 2006; SALAKHUTDINOV; HINTON, 2012). Essas técnicas baseiam-se no aprendizado hierárquico de características, similar ao processamento visual humano, por meio de abstrações em diferentes níveis (camadas) que auxiliam na extração de características.

Atualmente, algumas das principais técnicas de DL utilizadas são as Redes Neurais Convolucionais, do inglês *Convolutional Neural Networks* (CNNs) (LECUN et al., 1998), e Máquinas de Boltzmann Profundas, do inglês *Deep Boltzmann Machines* (DBMs) (SALAKHUTDINOV; HINTON, 2012). As CNNs são capazes de modelar as informações hierárquicas de maneira direta em basicamente três etapas. A primeira, corresponde à aplicação de convoluções no sinal de entrada e diferentes filtros, seguida por uma amostragem do sinal, e por fim um processo de normalização. Porém, sofrem com a necessidade de uma grande quantidade de dados rotulados para o processo de treinamento (supervisionado).

As DBMs possuem uma arquitetura específica com camadas ocultas compostas por Máquinas de Boltzmann, que se assemelha às Redes de Crença Profundas, do inglês *Deep Belief Networks* (DBNs) (HINTON; OSINDERO; TEH, 2006). Estas, por sua vez, são formadas pelo "empilhamento" de várias Máquinas de Boltzmann Restritas (*Restricted Boltzmann Machines* -

RBMs) (HINTON, 2002), ou seja, pode-se formar uma rede composta por diversas camadas constituídas de RBMs, em que a saída de cada uma é utilizada como entrada para uma outra, formando uma rede direcional. Por fim, as RBMs são classificadas como redes neurais estocásticas, onde um conjunto de neurônios ocultos (ou invisíveis) são responsáveis por modelar a distribuição de probabilidade dos dados de entrada, sem a necessidade de rótulos para o treinamento (sem supervisão). Essencialmente, uma RBM não é uma técnica de DL, uma vez que possui apenas uma camada de abstração (neurônios ocultos).

Com os avanços na área de visão computacional, as técnicas de DL passaram a ser empregadas não só para o domínio de imagens, mas também para o de vídeos, possibilitando a utilização de sistemas que atuam praticamente em tempo real, como no monitoramento por câmeras de segurança (MAHADEVAN et al., 2010; MOHAMMADI et al., 2016; AFONSO et al., 2018).

Entretanto, aplicações com vídeos aumentam expressivamente a complexidade e o custo computacional, uma vez que este domínio representa uma composição de diferentes imagens (*frames*) ao longo do tempo, elevando a quantidade de informações a serem tratadas e reconhecidas pelas técnicas de DL. Além disso, grande parte dos problemas deste domínio dizem respeito à classificação e reconhecimento de ações (FEICHTENHOFER; PINZ; ZISSERMAN, 2016; GOWDA, 2017; ULLAH et al., 2019), ao invés de eventos de alto nível ou eventos anômalos, tornando escassas as fontes de dados. Não obstante, grande parte das técnicas de DL tradicionais necessitam de dados rotulados para que a tarefa de classificação possa ser desempenhada.

1.1 Problema

As necessidades previamente mencionadas geram grandes dificuldades para os estudos desenvolvidos nesta área, uma vez que os eventos (de alto nível ou anômalos) estão relacionados a diferentes características nas imagens que compõe os vídeos, bem como a interação das entidades presentes nas cenas. Adicionalmente, na maioria dos casos, não há uma elevada disponibilidade de dados rotulados com os respectivos tipos de eventos ou anomalias, dificultando, e algumas vezes inviabilizando, o processo de aprendizado por técnicas de DL que demandam grandes quantidades de dados para a indução dos modelos.

Correlato à dificuldade apresentada, tem-se geralmente técnicas com arquiteturas complexas, que acentuam a dificuldade de um treinamento eficiente com poucos dados, prejudicando a capacidade de generalização para exemplos não conhecidos pela técnica de DL.

Diante disso, na tentativa de mitigar a ausência de bancos de dados para domínios específicos, pode-se fazer uso da Transferência de Aprendizado (TA), do inglês *Transfer Learning* (RAINA et al., 2007), que consiste no pré-treinamento das técnicas de DL em dados de um problema genérico e posterior transferência de parte do conhecimento para o problema

específico, por meio do processo de ajuste fino, do inglês *fine-tuning*, com dados do problema alvo (QUATTONI; COLLINS; DARRELL, 2008). Além disso, pode-se empregar a adaptação de domínio, do inglês *Domain Adaptation*, que consiste em mapear os dados de um domínio fonte para um domínio alvo utilizando uma função de mapeamento, permitindo o aproveitamento de carcterísticas de um domínio em outro (SUN; SHI; WU, 2015; LIU et al., 2019).

Em concordância com a adaptação de domínio, por exemplo, pode-se utilizar o treinamento semi-supervisionado, do inglês *semi-supervised training*, em que os dados sem classificação prévia são utilizados para treinar as técnicas de aprendizado profundo sob o paradigma não supervisionado e, posteriormente, os dados rotulados são utilizados no processo de ajuste fino dos parâmetros com a introdução das respectivas classes.

No que diz respeito à complexidade das arquiteturas, técnicas como as DBNs e DBMs podem ser utilizadas, uma vez que geralmente possuem menos camadas ocultas (usualmente 2 ou 3) que uma CNN, por exemplo (HINTON; OSINDERO; TEH, 2006; SALAKHUTDINOV; HINTON, 2012). Adicionalmente, com a utilização dessas técnicas, espera-se conseguir boas representações/características dos vídeos que simplifiquem a modelagem do problema.

Isto posto, este trabalho visa investigar o problema do treinamento de técnicas como RBMs, DBNs e DBMs para extrair características de vídeos com poucos dados rotulados, porém, com grandes volumes de *frames* para treinamento, empregando técnicas como adaptação de domínio e pré-treinamento de modelos baseados em energia.

Ademais, até onde se tem conhecimento, não há nenhum trabalho atualmente que aborde o problema de reconhecimento de eventos em vídeos por meio da abordagem e das técnicas citadas.

1.2 Objetivos

1.2.1 Objetivo Geral

Fazer uso de técnicas de aprendizado em profundidade, como RBMs, DBNs e DBMs, para a tarefa de reconhecimento de eventos em vídeos, apoiadas pela utilização da adaptação de domínio com dois paradigmas de treinamento, não supervisionado e supervisionado, possibilitando a extração de características de um domínio fonte (ações, por exemplo) para o alvo (eventos, por exemplo). Além disso, viabilizar métodos para o aproveitamento de bancos de dados de domínios diferentes, provendo uma biblioteca *open-source* com todos os modelos e técnicas implementadas nesse estudo para suprir a literatura de adaptação de domínio em vídeos e modelos baseados em energia.

1.2.2 Objetivos Específicos

Para alcançar os objetivos gerais, e gerar uma boa solução para o problema elucidado, os seguintes passos são considerados:

a) desenvolver um sistema de pré-processamento dos vídeos a fim de reduzir o número de *frames* necessários para o treinamento;

b) integrar a transferência de aprendizado possibilitada pela adaptação de domínio e pelos paradigmas de aprendizado utilizados;

c) testar e validar a metodologia proposta em bases de dados de referência (*benchmark*);e

d) possibilitar a reprodutibilidade do estudo realizado por meio de uma biblioteca de código aberto em Python para a comunidade contendo todos os estudos realizados.

1.3 Hipótese de Pesquisa

A principal hipótese dessa pesquisa é que os modelos profundos de redes neurais baseadas em energia, especificamente DBNs e DBMs, são capazes de atuar na tarefa de reconhecimento de eventos de alto nível a partir de ações, em vídeos extraídos de situações reais/cotidianas de elevada complexidade. Além disso, acredita-se que o processamento *Somatório* é capaz de reduzir o número de *frames* necessários para o treinamento das redes sem causar perdas significativas na desempenho preditiva dos modelos que o utilizam.

1.4 Estrutura da Dissertação

O restante da dissertação está organizada da seguinte maneira:

- O Capítulo 2 apresenta um panorama geral sobre o domínio dos vídeos;
- O Capítulo 3 apresenta conceitos e definições sobre o aprendizado em profundidade e as técnicas baseadas em energia utilizadas neste trabalho;
- O Capítulo 4 apresenta a abordagem proposta para mitigar os problemas levantados;
- O Capítulo 5 trata dos bancos de dados utilizados e da metodologia para a abordagem proposta;
- O Capítulo 6 mostra os resultados alcançados com a metodologia adotada;
- O Capítulo 7 apresenta a conclusão desta dissertação, bem como futuros trabalhos.
- O Capítulo 8 apresenta os trabalhos correlatos publicados e aceitos para publicação;

2 Domínio de Vídeos

Com o aumento expressivo de componentes eletrônicos que utilizam câmeras, sistemas de monitoramento e a elevada conectividade, a visão computacional tem sua atenção direcionada para aplicações apoiadas por vídeos.

Um vídeo é definido por uma sequência de imagens (*frames*), $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$, capturados em um intervalo de tempo t, em que a variação espaço-temporal dos objetos presentes nas imagens expressa a "movimentação" do mundo real. Adicionalmente, um vídeo pode conter efeitos de áudio. O conjunto de *frames* que representa o vídeo pode ser classificado de acordo com a complexidade das representações internas, bem como os níveis de interação entre as entidades no vídeo. A classificação é feita em 4 categorias: Atributos/Movimentos; Eventos/Ações de baixo nível; Interação; e Eventos de alto nível (JIANG et al., 2013).

Frente às categorias tem-se: a representação de mais baixo nível e/ou descrição de um *frame* como o Movimento, amplamente utilizada para reconhecimento de ações humanas, como a movimentação de membros do corpo, por exemplo (LIU; KUIPERS; SAVARESE, 2011). Já os Eventos/Ações de baixo nível representam uma determinada cadeia de movimentos, sendo esta realizada geralmente por uma entidade do *frame* (um carro ou uma pessoa, por exemplo). Quando essas ações são realizadas por mais de uma entidade, ou essas interagem entre si, é dada a categoria de Interação (JIANG et al., 2013).

Por fim, a categoria de mais alto nível, também chamada de Eventos complexos, representa a interação de entidades, ou uma sequência de ações, com longa duração temporal no vídeo. Um evento-exemplo pode ser uma festa de aniversário, composto por diversas ações e entidades. Isto posto, o reconhecimento de eventos pode ser compreendido como a detecção de localizações, temporais e espaciais, de um evento complexo na sequência de vídeos (JIANG et al., 2013).

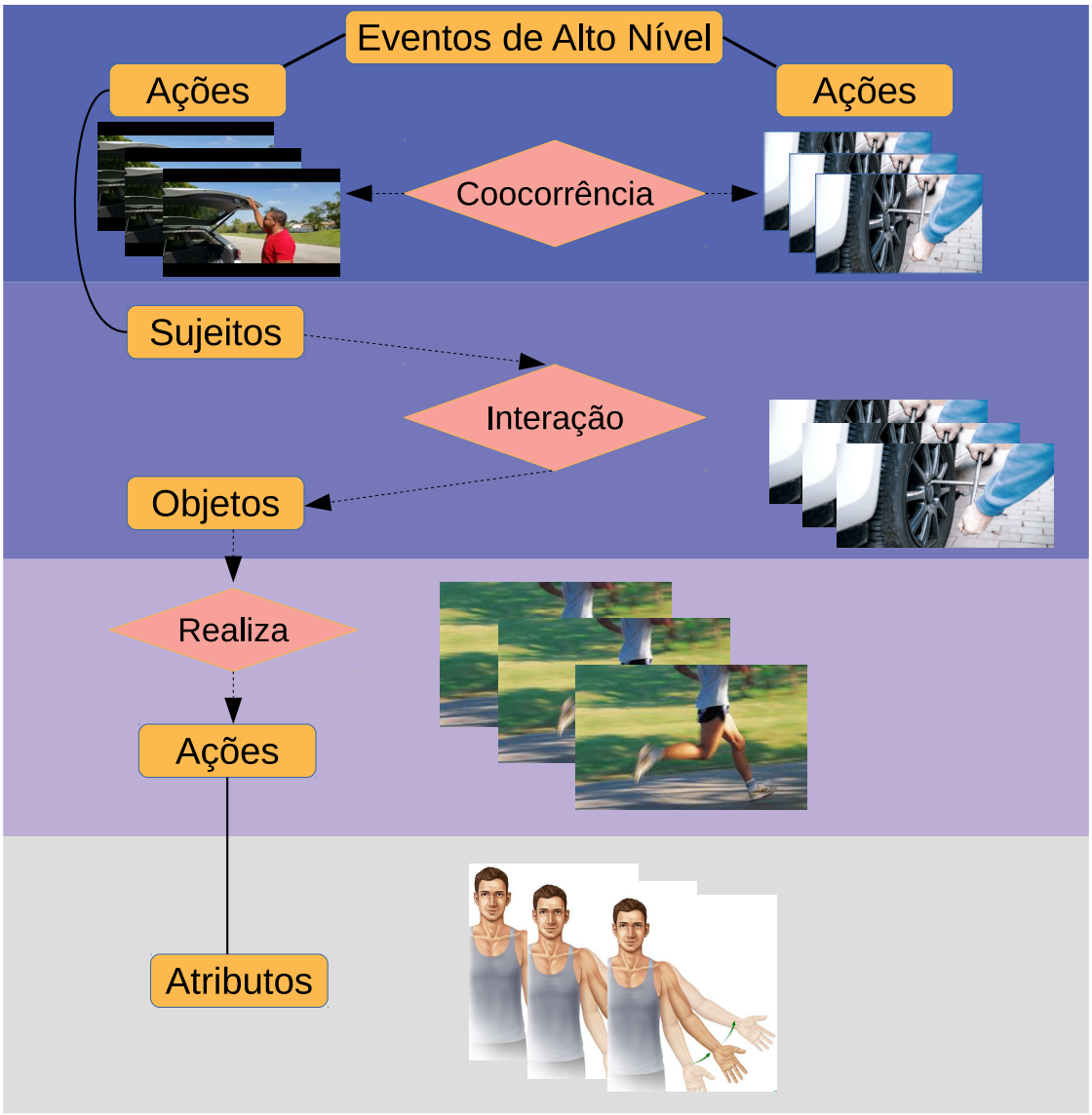
É importante ressaltar que na literatura não há um consenso ou uma padronização sobre a diferença efetiva entre ações e eventos, tornando-os intercambiáveis em grande parte das aplicações (BOBICK, 1997; JIANG et al., 2013). As categorias discutidas anteriormente são mostradas em sua forma hierárquica na Figura 1.

2.1 Trabalhos Relacionados

Uma vez abordados os conceitos primordiais sobre eventos em vídeos, nesta seção são mostrados os trabalhos mais próximos a esta dissertação, ou seja, realizam a detecção e/ou classificação de eventos complexos de alto nível com modelos baseados em grafos.

A utilização de algoritmos baseados em grafos possui duas categorias, os grafos

Figura 1 – Complexidade hierárquica no domínio de vídeos.



Fonte: Elaborado pelo autor.

direcionais e os não-direcionais. Exemplos sedimentados na literatura para o primeiro tipo incluem os baseados em Cadeias Ocultas de Markov (HMM, do inglês *Hidden Markov Models*) e Redes Bayesianas (BN, do inglês *Bayesian Networks*). Já para a segunda categoria tem-se os Campos Aleatórios de Markov (MRF, do inglês *Markov Random Fields*) e os Campos Aleatórios Condicionais (CRF, do inglês *Conditional Random Fields*).

Os modelos direcionais são responsáveis por modelar as dependências espaço-temporais em uma camada oculta de "neurônios", e as conexões direcionais representam as mudanças de estados da cadeia de Markov para o espaço de características observado. Li et al. (LI; ZHANG; LIU, 2008) abordaram, por exemplo, a modelagem de ações a partir de estados ocultos de HMMs, porém com a estimativa de poses salientes a partir de modelos de mistura gaussiana, para formar as ações de interesse.

O trabalho de Natarajan e Nevatia (NATARAJAN; NEVATIA, 2008) utilizou uma abordagem interessante para os modelos HMMs aplicados em reconhecimento de ações em pessoas, com o agrupamento de algumas cadeias de Markov para representar a composição de ações, enquanto camadas de HMMs foram responsáveis por representar as transições de poses dessas pessoas.

Já os trabalhos com grafos não-direcionais possuem foco nos modelos CRFs, com o início sendo marcado pelo trabalho de Vail et al. (VAIL; VELOSO; LAFFERTY, 2007) para reconhecimento de ações. Os autores mostraram que os CRFs podem ser empregados em uma abordagem discriminativa, levando em consideração toda a sequência de observações, e não mais observações individuais. O sucesso da abordagem se deu por uma característica do algoritmo, o aprendizado das probabilidades condicionais dos estados, que se correlacionam com a sequência espaço-temporal das observações.

Ainda neste âmbito, Conolly (CONNOLLY, 2007) empregou CRFs para a modelagem e reconhecimento de eventos em imagens de câmeras de segurança. Mesmo atingindo bons resultados para a época, o autor ressalta que os atributos descritores das poses das pessoas devem ser cuidadosamente analisados e modelados, se possível.

Porém, a técnica possui suas limitações, principalmente quando há o relacionamento de causa e efeito na sequência de *frames* analisada. Além disso, é importante ressaltar que todas necessitam de dados pré-processados, ou seja, não são capazes de utilizar os dados brutos provenientes de câmeras, por exemplo.

Nos anos seguintes surgiram diversos bancos de dados maiores, porém, focados em subsidiar a área de classificação de ações de baixo e alto nível, mas não de eventos complexos. Por consequência, os trabalhos desenvolvidos atualmente abordam a classificação de ações, e majoritariamente, a aplicação de redes neurais convolucionais com variações que empregam descritores de características temporais, como o fluxo óptico, compondo a entrada da rede (SIMONYAN; ZISSERMAN, 2014), e variações que agregam diferentes arquiteturas para extrair informações espaço-temporais (ZHU et al., 2018).

Estes fatos motivam o estudo de modelos baseados em grafos não direcionais como uma opção para tratar essa "deficiência", como os modelos de aprendizado em profundidade baseados em energia (RBMs, DBNs e DBMs).

2.2 Bancos de Dados Públicos

No domínio de vídeos, alguns bancos de dados ganharam destaque nos últimos anos, principalmente pela dificuldade de se obter vídeos com suas respectivas classificações, ao passo que sejam representativos dos problemas do mundo real. Neste âmbito, a Tabela 1 apresenta os principais bancos de dados utilizados e empregados na literatura nos domínios de ações e

eventos de alto nível.

Banco	#Vídeos	#Classes	Ano	Fundo da Cena	Domínio
KTH	600	6	2004	Estático	Ação
Weizmann	81	9	2005	Estático	Ação
Kodak	1.358	25	2007	Dinâmico	Ação (anotação)
Hollywood2	1.707	12	2009	Dinâmico	Ação
Olympic Sports	800	16	2010	Dinâmico	Ação
UCSD	98	2	2010	Dinâmico	Evento (anomalia)
HMDB51	6.766	51	2011	Dinâmico	Ação/Evento
CCV	9.317	20	2011	Dinâmico	Ação
UCF-101	13.320	101	2012	Dinâmico	Ação/Evento
THUMOS-2014	18.394	101	2014	Dinâmico	Ação
Sports-1M	1.133.158	487	2014	Dinâmico	Ação
Kinetics-600	495.547	600	2018	Dinâmico	Ação

Tabela 1 – Principais bancos de dados.

Em 2004 surgiu o banco de dados KTH (SCHULDT; LAPTEV; CAPUTO, 2004), um dos mais antigos no que diz respeito ao reconhecimento de ações humanas em vídeos. É relativamente pequeno e possui poucas ações. O fundo das cenas (*background*) é controlado, sem grandes variações que possam atrapalhar a classificação das ações. No ano seguinte surgiu o banco *Weizmann* (GORELICK et al., 2007). É um conjunto de vídeos relativamente pequeno, possui 81 gravações e 9 classes de ações humanas. Além disso, o fundo das cenas é estático e possui poucas variações.

Já em 2007, surgiu o banco *Kodak* (LOUI et al., 2007), que sumariza conceitos e ações de diversos consumidores da *Kodak* segundo a ontologia da empresa. O banco possui um grande número de vídeos e ações, ao passo que o fundo das cenas é dinâmico.

Em 2009, o banco de dados *Hollywood* foi incrementado por (MARSZALEK; LAPTEV; SCHMID, 2009) para o chamado *Hollywood2*, composto por 1.707 vídeos. Estes são compostos por 12 ações retiradas de 69 filmes de *Hollywood*, os quais representam grande dificuldade para o tratamento devido ao fundo das cenas ser altamente variável, e com grande movimentação de câmeras.

No ano seguinte, surgiu o banco *Olympic Sports* (NIEBLES; CHEN; FEI-FEI, 2010), composto por 800 vídeos e 16 ações humanas no contexto de esportes olímpicos. A diferença deste diz respeito à origem dos vídeos, sendo todos baixados da internet para compor o banco.

Ainda em 2010, surgiu um importante banco para detecção de eventos anômalos em ambientes abertos de uma universidade, o UCSD (MAHADEVAN et al., 2010). Este é divido em dois *subsets* (subgrupos) que representam duas cenas, o Peds1 e o Peds2. O primeiro, representa pedestres caminhando a favor ou contra a câmera, enquanto o Peds2 captura a movimentação de pessoas no plano paralelo à câmera. Por ser um banco de detecção de eventos anômalos, possui duas classes.

Já em 2011, emergiu o banco HMDB51 (KUEHNE et al., 2011), com 6.766 vídeos que compreendem um total de 51 ações de pessoas. Os vídeos são provenientes de várias fontes da internet, incluindo filmes e vídeos de usuários do YouTube. Novamente, dada a composição heterogênea das fontes, os clipes possuem o fundo das cenas dinâmico e altamente variável. É categorizado como um banco de classificação de ações, porém esta categorização pode ser intercambiável devido ao agrupamento de ações em macro-classes que o banco possui. No mesmo ano, foi publicado o banco *Columbia Consumer Videos* (CCV) (JIANG et al., 2011), com um total de 9.317 vídeos. Este tem a particularidade de ser composto apenas por vídeos feitos por usuários do YouTube. Os vídeos possuem 20 classes correspondentes que incluem simples objetos, cenas/vistas naturais, eventos relacionados a esportes e atividades de convívio social.

Em 2012, Soomro e seus colaboradores lançaram o UCF101 (SOOMRO; ZAMIR; SHAH, 2012), um dos mais importantes e desafiadores banco de dados de vídeos da literatura, sendo o sucessor dos bancos UCF11 e UCF50. Trata-se de um conjunto de dados desafiador devido à elevada variabilidade intra-classes e interclasses, bem como as variações nas filmagens e no *background* das cenas. É composto por 13.320 vídeos de usuários do YouTube, e possui a categorização de 101 ações de pessoas em diversos aspectos de interação entre as entidades presentes nas cenas. As 101 ações são categorizadas pelos autores em 5 grandes classes, que agrupam os vídeos em características de alto nível como a prática esportiva ou a interação entre pessoas e objetos.

Dois anos após a criação do UCF101, Jiang e seus colaboradores incrementaram o banco para a competição THUMUS, gerando o THUMOS-2014 (JIANG et al., 2014). O novo banco possui as mesmas características, porém, conta com 18.394 vídeos, 5.074 a mais que o UCF101. Ainda em 2014, Karpathy et al. agruparam 1.133.158 vídeos sobre esportes diretamente do YouTube no chamado Sports-1M (KARPATHY et al., 2014), formando o maior banco de dados do gênero até então. O banco conta com 487 ações humanas relacionadas à pratica esportiva. Novamente, por serem vídeos retirados da internet, não contam com controle de *background* nas cenas.

Por fim, um dos últimos bancos de dados criados na área é o Kinetics-600, proposto por (CARREIRA et al., 2018) como uma extensão do Kinetics-400. O banco é composto por 495.547 videoclipes distribuídos em 600 classes de ações humanas extraídas de vídeos do YouTube, como os supracitados. Porém, como mencionado, o banco aborda apenas ações humanas.

3 Aprendizado em Profundidade

No decorrer do desenvolvimento tecnológico, surgiu a necessidade de métodos automáticos para a análise de dados, fato que impulsionou o desenvolvimento da área de aprendizado de máquina (AM). Esta, faz uso de algoritmos capazes de se adaptarem independentemente, que buscam aprender padrões ocultos contidos nos dados. Este aspecto fez as técnicas de AM emergirem de maneira expressiva, gerando resultados promissores (BISHOP, 1995).

Porém, as aplicações no mundo real têm se tornado cada vez mais complexas, seja pela diversidade de dispositivos que geram dados ou pela própria natureza destes, que passam a construir relacionamentos intrínsecos difíceis de serem modelados. Um exemplo de complexidade diz respeito à análise de imagens RGB, em que cada componente de cor (*red*, *green*, *blue*) representa um plano bidimensional que compõe a imagem $2D$, bem como as mais diversas aplicações que tangem desde a medicina até ambientes industriais complexos.

Quando analisamos imagens, a luminosidade do ambiente e fatores como rotação destas (mesmo que em pequenos ângulos) podem fazer com que a intensidade das cores dos *pixels* sofram variações de grande magnitude, mesmo que ao olho humano isso seja pouco perceptível. Deste modo, levantou-se a questão de como extrair características relevantes de um domínio altamente variável (BISHOP, 1995). A partir da necessidade de embutir o próprio processo de caracterização dos dados às técnicas de AM, emergiu uma sub-área dentro do aprendizado de máquina, denominada aprendizado em profundidade ou profundo (*Deep Learning*). As técnicas de DL possuem a particularidade de utilizar os dados brutos como entrada, no caso de imagens, os *pixels*, para extrair informações relevantes sem necessariamente utilizar algum pré-processamento.

Em suma, busca-se representar informações sofisticadas a partir de representações mais simples, ao passo que relacionamentos complexos são modelados de maneira mais fácil. Essa abordagem se aproxima ao aprendizado do ser humano, em que o cérebro torna as informações hierárquicas a partir de entradas simples, e é capaz de construir conhecimentos robustos e complexos.

Historicamente falando, o aprendizado em profundidade parece ser uma área de estudo recente, porém, o início do seu desenvolvimento se deu próximo da década de 1940, dadas as necessidades da segunda guerra mundial. Obviamente, nos primórdios do seu desenvolvimento, a tecnologia era mais limitada e as aplicações bem específicas, e em sua maioria para uso militar. Desde o início, os algoritmos objetivaram modelar computacionalmente o aprendizado biológico humano, ou seja, se aproximar do processo de aprendizado do cérebro humano. Um dos principais exemplos são as Redes Neurais Artificiais, do inglês *Artificial Neural Networks*, com capacidade de aplicação em diversos domínios (BISHOP, 1995).

Em 1943, foi concebido o neurônio de McCulloch-Pitts (MCCULLOCH; PITTS, 1943), o primeiro modelo de um neurônio computacional, linear, que tenta imitar o comportamento biológico. Esse modelo era capaz de distinguir duas categorias, uma positiva e outra negativa, por meio da correta seleção de suas conexões neurais (pesos sinápticos). Já na década de 1950 surgiu o modelo Perceptron (ROSENBLATT, 1958), capaz de aprender os pesos de maneira automática a partir dos dados de entrada. Entretanto, este modelo ainda era limitado, sendo famoso por não ser capaz de aprender a função lógica XOR, o que causou críticas e perda de popularidade da técnica.

Entre 1950 e 1980 houve um certo hiato no desenvolvimento das técnicas de inteligência artificial, que foi quebrado com o movimento chamado conexionismo (RUMELHART et al., 1988; MCCLELLAND; RUMELHART; HINTON, 1986). Em suma, o movimento pregou que a utilização de diversos neurônios artificiais, de maneira concomitante, é capaz de gerar um comportamento inteligente. Em 1986, Geoffrey Hinton (HINTON, 1986) introduziu o conceito de representação distribuída, presente até a atualidade no aprendizado profundo. Este, nos fala que cada entrada do sistema deve ser representada por várias características, e estas por sua vez, devem representar o máximo de dados de entrada possível.

Ainda no movimento conexionista, houve o desenvolvimento e a utilização do algoritmo de retro-propagação (*back-propagation*) (RUMELHART; HINTON; WILLIAMS, 1986; LECUN, 1987) para o treinamento de redes neurais. Esse algoritmo funciona calculando os gradientes de uma função de perda (*loss function*), que guia o processo de aprendizado para o ajuste dos pesos sinápticos. Este foi um fato marcante, uma vez que possibilitou diversos avanços e é amplamente utilizado até hoje, para as mais diferentes redes neurais artificiais.

Finalmente, em 2006, Geoffrey Hinton deu início à terceira "revolução" no aprendizado em profundidade, agora sendo efetivamente profundo no que diz respeito à profundidade (quantidade de camadas neurais) das redes neurais artificiais, introduzindo a chamada *Deep Belief Network* (DBN), ou Rede de Crença em Profundidade (HINTON; OSINDERO; TEH, 2006). Esse fato, popularizou o termo "aprendizado em profundidade", uma vez que possibilitou o treinamento de redes com várias camadas, e elevou o patamar dentro da inteligência artificial.

O sucesso das DBNs também foi responsável por popularizar os chamados modelos baseados em energia, que abrangem principalmente as *Restricted Boltzmann Machines* (RBMs), as *Deep Belief Networks* e *Deep Boltzmann Machines* (DBMs). Estes foram e são aplicados em diversos problemas: Salakhutdinov, Mnih e Hinton (2007), Larochelle e Bengio (2008), Salakhutdinov e Hinton (2009), Salakhutdinov e Hinton (2012), Khojasteh et al. (2019), Passos e Papa (2018), Passos et al. (2019).

3.1 Modelos Baseados em Energia

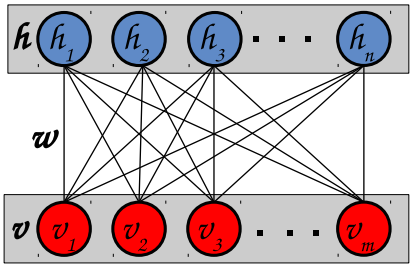
Nesta seção são apresentados três modelos baseados em energia, amplamente utilizados em diversas tarefas, tais como pré-treinamento de redes profundas, classificação de padrões, extração de características e geração de dados.

3.1.1 Máquinas de Boltzmann Restritas

As Máquinas de Boltzmann Restritas são redes neurais estocásticas baseadas em princípios físicos de energia, entropia e temperatura. São compostas basicamente por duas camadas de neurônios/unidades (visíveis e ocultos), capazes de modelar problemas sob os paradigmas de aprendizado não supervisionado (HINTON, 2002) ou supervisionado (LAROCHELLE; BENGIO, 2008). A RBM é uma variação da clássica Máquina de Boltzmann (ACKLEY; HINTON; SEJNOWSKI, 1988), porém, possui restrições de conexão entre os neurônios da mesma camada.

Em suma, uma RBM é a representação de um grafo bipartido, com conexões não direcionais. A Figura 2 descreve a arquitetura de uma Máquina de Boltzmann Restrita, com a camada visível **v** possuindo m unidades, e a camada oculta **h** com n neurônios. A matriz **w**, com valores reais, modela os pesos (conexão neural) entre os neurônios visíveis e ocultos, possuindo dimensão $m \times n$.

Figura 2 – Arquitetura de uma RBM.



Fonte: Elaborado pelo autor.

Inicialmente, as RBMs foram desenvolvidas usando neurônios visíveis e ocultos com estados binários, as chamadas Máquinas de Boltzmann Restritas de Bernoulli[1], do inglês *Bernoulli RBMs* (BRBMs), cujos estados das unidades são amostrados a partir da distribuição de Bernoulli. Posteriormente, Welling et al. (WELLING; ROSEN-ZVI; HINTON, 2005) e Hinton (HINTON, 2012) apresentaram variações para as unidades que podem ser usadas em uma RBM, como as binomiais, as unidades lineares retificadas (ReLU) e as gaussianas. As variações mencionadas são generalizações da BRBM, portanto, os conceitos relacionados a esta são apresentados.

[1] Frequentemente o nome Bernoulli é omitido por ser o modelo base de RBMs, tornando as siglas BRBM e RBM intercambiáveis

Sejam **v** e **h** as unidades visíveis e ocultas binárias, respectivamente, ou seja, $\mathbf{v} \in \{0,1\}^m$ e $\mathbf{h} \in \{0,1\}^n$. A energia de uma Máquina de Boltzmann Restrita de Bernoulli é modelada como segue:

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i h_j w_{ij}, \tag{1}$$

em que **a** e **b** são os valores dos vieses (*biases*) das unidades visíveis e ocultas, respectivamente. A probabilidade de uma configuração conjunta $(\mathbf{v},\mathbf{h})$ é calculada como segue:

$$P(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}, \tag{2}$$

onde o denominador da equação é um fator de normalização conhecido como função de partição, que leva em conta todas as possíveis configurações envolvendo unidades visíveis e ocultas, sendo intratável para espaços de alta dimensão como imagens, por exemplo.

Em suma, o processo de treinamento de uma BRBM tem por objetivo maximizar as probabilidades observadas de uma configuração $P(\mathbf{v})$, ao passo que é necessário estimar e ajustar os valores de **w**, **a** e **b**. Portanto, a próxima seção descreve esse procedimento.

3.1.1.1 O Treinamento

Por se tratar de um problema de otimização, os parâmetros da BRBM podem ser otimizados através da técnica da subida do gradiente estocástico, método dual para a descida do gradiente estocástico (*Stochastic Gradient Descent*- SGD), aplicado no logaritmo da verossimilhança (*Log-Likelihood*) dos dados de treinamento. A verossimilhança é calculada para uma amostra apresentada às unidades visíveis, e sua probabilidade é obtida como segue:

$$P(\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}. \tag{3}$$

Isto posto, para atualizar os pesos e vieses, é necessário computar as seguintes derivadas:

$$\frac{\partial \log P(\mathbf{v})}{\partial w_{ij}} = E[h_j v_i]_{data} - E[h_j v_i]_{modelo}, \tag{4}$$

$$\frac{\partial \log P(\mathbf{v})}{\partial a_i} = E[v_i]_{data} - E[v_i]_{modelo}, \tag{5}$$

$$\frac{\partial \log P(\mathbf{v})}{\partial b_j} = E[h_j]_{data} - E[h_j]_{modelo}, \tag{6}$$

em que $E[\cdot]$ representa a esperança estatística sob uma distribuição, e $E[\cdot]_{data}$ e $E[\cdot]_{modelo}$ representam as probabilidades dos dados originais e reconstruídos, respectivamente.

Em termos práticos, podemos computar $E[h_j v_i]_{data}$ considerando **h** e **v** como segue:

$$E[\mathbf{hv}]_{data} = P(\mathbf{h}|\mathbf{v})\mathbf{v}^T, \tag{7}$$

em que $P(\mathbf{h}|\mathbf{v})$ é a probabilidade associada ao espaço latente, representado pelas unidades ocultas **h**, dada uma observação no espaço visível **v** (dado de treinamento):

$$P(h_j = 1|\mathbf{v}) = \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + b_j\right), \tag{8}$$

onde $\sigma(\cdot)$ é a função sigmoide-logística[2]. Deste modo, $E[\mathbf{hv}]_{data}$ é obtida da seguinte maneira: dada uma amostra de treinamento $\mathbf{x} \in X$, onde X é um conjunto de treinamento, precisa-se ajustar $\mathbf{v} \leftarrow \mathbf{x}$, e então utilizar a Equação 8 para obter $P(\mathbf{h}|\mathbf{v})$ e, através da Equação 7, obtém-se $E[\mathbf{hv}]_{data}$.

Uma vez obtidas as estatísticas a partir dos dados ($E[\mathbf{hv}]_{data}$), o próximo passo é obter $E[\mathbf{hv}]_{modelo}$, o qual representa a distribuição aprendida pelo modelo. Da mecânica estatística, uma estratégia é obter a esperança do modelo por meio da técnica de amostragem de Gibbs, um algoritmo baseado no método de Monte Carlo com Cadeias de Markov (MCMC, do inglês *Markov Chain Monte Carlo method*). Esta inicia as unidades visíveis com valores aleatórios, e atualiza as unidades ocultas utilizando a Equação 8, seguida pela atualização das unidades visíveis usando $P(\mathbf{v}|\mathbf{h})$, dada por:

$$P(v_i = 1|\mathbf{h}) = \sigma\left(\sum_{j=1}^{n} w_{ij} h_j + a_i\right), \tag{9}$$

e então, atualizam-se novamente as unidades ocultas utilizando a Equação 8 até um critério de convergência da cadeia ser atingido, como k iterações, por exemplo. Em suma, a técnica possibilita obter uma estimativa de $E[\mathbf{hv}]_{modelo}$ a partir de valores aleatórios.

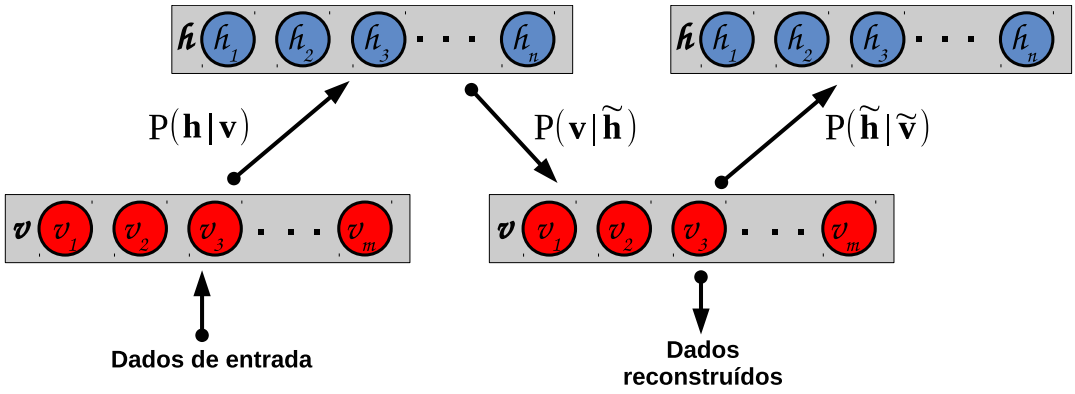
Entretanto, tal procedimento é computacionalmente custoso, e inviável para espaços de alta dimensão, uma vez que a convergência da cadeia é garantida quando as iterações tendem ao infinito $k \to +\infty$. Por conta desta dificuldade, alguns trabalhos apresentaram alternativas à amostragem de Gibbs, sendo a principal delas a Divergência Contrastiva (*Contrastive Divergence* - CD) (HINTON, 2002).

[2] A função sigmoide-logística é computada pela seguinte equação: $\sigma(x) = 1/(1 + \exp(-x))$.

3.1.1.2 Divergência Contrastiva

Hinton (HINTON, 2002) introduziu uma metodologia mais simples, eficaz e rápida para o cálculo de $E[\mathbf{hv}]_{modelo}$ tendo como base a ideia de divergência contrastiva. Em suma, a simplicidade dá-se pelo fato da inicialização das unidades visíveis com amostras dos dados de treinamento, para inferir os estados latentes utilizando as probabilidade da Equação 8. Uma vez calculadas estas probabilidades pode-se calcular os estados das unidades visíveis, ou seja, a reconstrução dos dados, por meio da Equação 9. Este procedimento é equivalente a execução da amostragem de Gibbs usando $k = 1$, com os valores da cadeia inicializados com amostras de treinamento. A Figura 3 ilustra essa abordagem.

Figura 3 – Ilustração do método baseado em divergência contrastiva.



Fonte: Elaborado pelo autor.

O procedimento apresentado na Figura 3 pode ser iterado por k vezes, porém, na maioria das aplicações $k = 1$ é suficiente para gerar boas aproximações, se, e somente se, a cadeia possui uma alta taxa de mistura, isto é, consegue convergir rapidamente ao passo que reduz as dependências anteriores.

Diante do exposto anteriormente, pode-se calcular $E[\mathbf{hv}]_{modelo}$ como segue:

$$E[\mathbf{hv}]_{modelo} = P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T. \tag{10}$$

Portanto, o problema de obtenção das estatísticas do modelo, e consequentemente do aprendizado de seus parâmetros, é sanado. Isso possibilita atualizar a matriz de pesos $\mathbf{W}$, como segue:

$$\begin{aligned} \mathbf{w}^{t+1} &= \mathbf{w}^t + \eta(E[\mathbf{hv}]_{data} - E[\mathbf{hv}]_{modelo}) \\ &= \mathbf{w}^t + \eta(P(\mathbf{h}|\mathbf{v})\mathbf{v}^T - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T), \end{aligned} \tag{11}$$

onde $\mathbf{w}^t$ é a matriz de pesos no instante t, e η corresponde à taxa de aprendizado. Adicionalmente, os vieses das unidades visíveis e ocultas são atualizados seguindo as formulações:

$$\begin{aligned} \mathbf{a}^{t+1} &= \mathbf{a}^t + \eta(\mathbf{v} - E[\mathbf{v}]_{modelo}) \\ &= \mathbf{a}^t + \eta(\mathbf{v} - \tilde{\mathbf{v}}), \end{aligned} \tag{12}$$

e

$$\begin{aligned} \mathbf{b}^{t+1} &= \mathbf{b}^t + \eta(E[\mathbf{h}]_{data} - E[\mathbf{h}]_{modelo}) \\ &= \mathbf{b}^t + \eta(P(\mathbf{h}|\mathbf{v}) - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})), \end{aligned} \tag{13}$$

em que $\mathbf{a}^t$ e $\mathbf{b}^t$ são os valores de *bias* das unidades visíveis e ocultas no momento t, respectivamente. Em resumo, com as Equações 11, 12 e 13 pode-se atualizar os parâmetros da RBM.

3.1.1.3 RBMs para Dados Contínuos

Além de neurônios binários, as RBMs podem acomodar neurônios visíveis capazes de trabalhar com dados não binários, ou seja, contínuos, úteis para modelar diferentes tipos de entradas ou sinais que variam em largas faixas de valores. As mudanças que possibilitam esta utilização ocorrem na função de energia, como segue:

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i=1}^{m} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{v_i}{\sigma_i} h_j w_{ij}, \tag{14}$$

em que σ_i representa o desvio padrão dos dados de entrada, e σ_i^2 a variância, para cada neurônio i. Considerando as derivadas, do mesmo modo que em uma RBM binária, as probabilidades condicionais dos neurônios da camada visível se tornam:

$$P(v_i = 1|\mathbf{h}) \sim \mathcal{N}\left(\sum_{j=1}^{n} w_{ij} h_j + a_i, \sigma_i^2\right). \tag{15}$$

É fácil notar que quando são apresentados dados com média zero e desvio padrão unitário ($\sigma_i = 1$), ou seja, dados normalizados em uma Gaussiana padrão, a Equação 15 torna-se simples e fácil de ser empregada, uma vez que o procedimento de aprendizado mantém-se o mesmo.

3.1.2 Redes de Crença em Profundidade

Com o aumento da complexidade dos problemas, surgiu a necessidade de melhorar a representação dos dados, e aprender características profundas e intrínsecas do domínio. Esse fato fomentou o desenvolvimento das Redes de Crença em Profundidade (*Deep Belief Networks*

- DBNs), compostas por uma ou mais RBMs empilhadas que formam uma rede híbrida com conexões direcionais após o treinamento.

A intuição das DBNs é tal que, o empilhamento de RBMs auxilia no processo de extração das características de alto nível, isto é, são capazes de modelar a hierarquia dos dados apresentados à primeira camada visível. Com isso, espera-se que no decorrer das passagens pelas camadas intermediárias, a função *Log-Likelihood* tenha seu limite assintótico (*lower bound*) aumentado, representando uma melhor modelagem da distribuição dos dados.

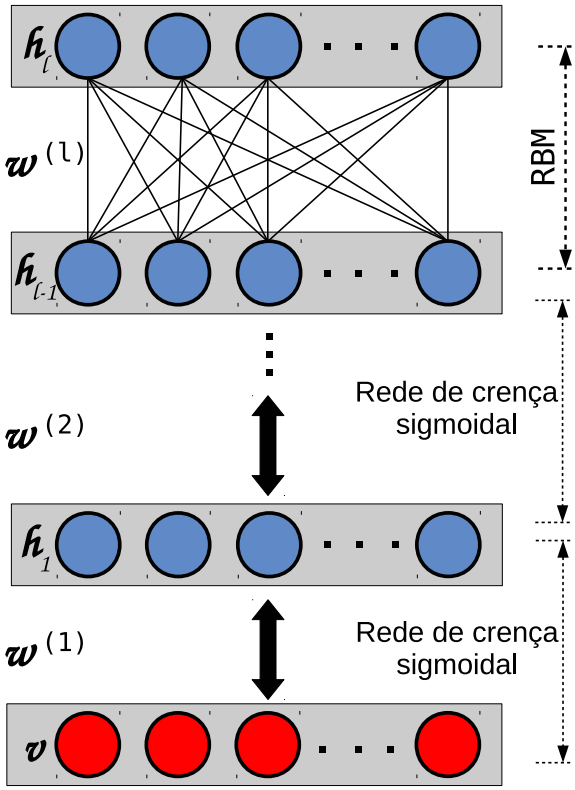
O treinamento de uma DBN é relativamente simples, cada RBM da arquitetura é treinada "isoladamente", ou seja, a rede final é formada bloco a bloco. Esse procedimento é chamado de *Greedy Layer-wise training*, e utiliza o processo de aprendizado apresentado na Seção 3.1.1.1, cuja RBM que está sendo treinada em uma camada arbitrária não considera outras durante o seu treinamento. Após o treinamento de um bloco, os anteriores tornam-se redes de crença sigmoidais, uma vez que suas ativações são ditadas pela função logística, além da possibilidade de realizar tanto geração quanto inferência de dados.

A Figura 4 ilustra a arquitetura de uma DBN, em que cada RBM de uma dada camada é representada conforme ilustrado na Figura 2, novamente com os respectivos vieses ocultados. Nesse caso, temos uma DBN composta por l camadas, sendo $\mathbf{w}^l$ as conexões neurais entre as RBMs da camada l. É importante ressaltar que as unidades ocultas da camada l tornam-se as visíveis da $l+1$.

Após a etapa previamente descrita, também conhecida como pré-treinamento, Hinton et al. (HINTON; OSINDERO; TEH, 2006) propõe a realização de um ajuste adicional e final dos parâmetros da rede (*fine-tuning*), o qual pode ser realizado sob os paradigmas não supervisionado ou supervisionado. O primeiro é realizado pelo algoritmo *Wake-Sleep* (HINTON et al., 1995), cuja intuição é apresentar os dados à camada de entrada $\mathbf{v}$, propagar o sinal pela rede e, atualizar as conexões de acordo com o resultado obtido. Posteriormente, é realizada a propagação "para baixo", que representa a geração de amostras para os dados iniciais, seguida pelo ajuste dos pesos frente a geração obtida.

Já o ajuste supervisionado utiliza o algoritmo de Retro-propagação ou Gradiente Descendente, a fim de ajustar as matrizes de peso $\mathbf{w}^l$, $l = 1, 2, \ldots, l$. O algoritmo de otimização trabalha minimizando uma medida de erro, geralmente o de classificação, a partir da saída de uma camada extra adicionada ao topo da DBN após o *Greedy Layer-wise training*. Essa camada geralmente é composta por unidades logísticas ou do tipo *softmax*, podendo também ser substituída por uma técnica de classificação de padrões supervisionada, como máquinas de vetores de suporte (CORTES; VAPNIK, 1995).

Figura 4 – Arquitetura padrão de uma DBN.



Fonte: Elaborado pelo autor.

3.1.3 Máquinas de Boltzmann em Profundidade

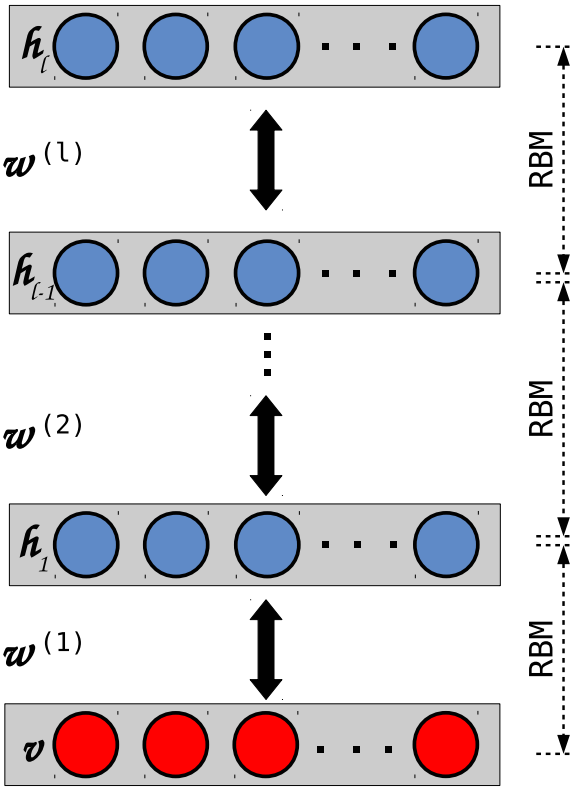
Dado o sucesso dos modelos baseados em energia, Salakhutdinov e Hinton (SALAKHUTDINOV; HINTON, 2009) desenvolveram as *Deep Boltzmann Machines*, outra variação proveniente do processo de empilhamento de RBMs, porém, com algumas diferenças das DBNs. A Figura 5 apresenta a arquitetura de uma DBM, generalizada para L camadas, em que cada RBM do empilhamento é representada como mostrado na Figura 2.

Seja uma DBM com duas camadas, em que $\mathbf{h}_1$ e $\mathbf{h}_2$ são as unidades ocultas da primeira e da segunda camada, respectivamente, é possível definir a energia global pela Equação 16:

$$E(\mathbf{v}, \mathbf{h}_1, \mathbf{h}_2) = -\sum_{i=1}^{m_1}\sum_{j=1}^{n_1} v_i h_{1j} w_{ij}^1 - \sum_{i=1}^{m_2}\sum_{j=1}^{n_2} h_{1i} h_{2j} w_{ij}^2, \tag{16}$$

onde m_1 e m_2 são as quantidades de unidades visíveis na primeira e segunda camada, respectivamente, e n_1 e n_2 correspondem às quantidades de unidades ocultas na primeira e segunda camada, respectivamente. Adicionalmente, as matrizes de pesos $\mathbf{w}^1_{m_1 \times n_1}$ e $\mathbf{w}^2_{m_2 \times n_2}$ são as conexões entre os vetores $\mathbf{v}$ e $\mathbf{h}_1$, e os vetores $\mathbf{h}_1$ e $\mathbf{h}_2$, respectivamente. É importante ressaltar que para simplificar, sem gerar perdas na interpretação, os termos referentes aos vieses foram omitidos. A probabilidade mínima que o modelo atribui a um vetor de entrada $\mathbf{v}$ é dada como

Figura 5 – Arquitetura padrão de uma DBM.



Fonte: Elaborado pelo autor.

segue:

$$P(\mathbf{v}) = \frac{1}{Z} \sum_{\mathbf{h}_1, \mathbf{h}_2} e^{-E(\mathbf{v}, \mathbf{h}_1, \mathbf{h}_2)}. \tag{17}$$

em que Z representa a função de partição (HINTON, 2012).

As probabilidades condicionais sobre as unidades visíveis e sobre ambos espaços latentes são dadas pelas Equações 18, 19 e 20:

$$P(v_i = 1|\mathbf{h}_1) = \sigma\left(\sum_{j=1}^{n_1} w_{ij}^1 h_{1j}\right), \tag{18}$$

$$P(h_{2j} = 1|\mathbf{h}_1) = \sigma\left(\sum_{i=1}^{m_2} w_{ij}^2 h_{1i}\right), \tag{19}$$

e

$$P(h_{1j} = 1|\mathbf{v}, \mathbf{h}_2) = \sigma\left(\sum_{i=1}^{m_1} w_{ij}^1 v_i + \sum_{j=1}^{n_2} w_{ij}^2 h_{2j}\right). \tag{20}$$

Entre os anos de 2010 e 2012, Salakhutdinov e Hinton (2012) estudaram e propuseram um método eficiente para pré-treinar as DBMs, alcançando o estado da arte para a classificação de dígitos manuscritos na época.

O pré-treinamento proposto utiliza conceitos do treinamento de uma DBN, ou seja, as RBMs são treinadas camada a camada utilizando o método da Divergência Contrastiva. Porém, os pesos de cada RBM possuem seus valores dobrados quando utilizados para amostrar a camada oculta, e mantidos em seus valores originais para a amostragem da camada visível. Este procedimento tenta suprir a falta das conexões de uma camada superior que compõe a respectiva DBM. Este comportamento é mantido até que a última RBM seja adicionada à pilha, invertendo o processo de escala, ou seja, as conexões para a amostragem da última camada oculta têm seus valores mantidos, ao passo que para a amostragem da camada inferior (visível) seus valores são dobrados (SALAKHUTDINOV; HINTON, 2012).

Outra particularidade das DBMs está no processo de atualização de seus parâmetros, especificamente no cálculo da esperança estatística proveniente dos dados (dados de entrada de cada camada oculta, não aos dados da primeira camada visível). Este processo emprega o conceito de campo-médio (do inglês, *mean-field*[3]), em que todas as probabilidades condicionais de uma dada camada oculta são atualizadas iterativamente a partir da fixação das demais (como mostrado nas equações supracitadas). Uma vez aproximadas as condicionais, a esperança do modelo pode ser calculada normalmente com a amostragem de Gibbs, através da Divergência Contrastiva (SALAKHUTDINOV; HINTON, 2012).

Uma vez que o aprendizado é realizado pelo método da Divergência Contrastiva, o modelo generativo pode ser escrito pela Equação 21:

$$P(\mathbf{v}) = \sum_{\mathbf{h}_1} P(\mathbf{h}_1) P(\mathbf{v}|\mathbf{h}_1), \tag{21}$$

em que $P(\mathbf{h}_1) = \sum_{\mathbf{v}} P(\mathbf{h}_1, \mathbf{v})$. Consequentemente, o processo continua para a segunda RBM, substituindo $P(\mathbf{h}_1)$ por $P(\mathbf{h}_1) = \sum_{\mathbf{h}_2} P(\mathbf{h}_1, \mathbf{h}_2)$ (SALAKHUTDINOV; HINTON, 2012).

[3] Os autores mostraram que 25 iterações são suficientes para uma convergência aceitável

4 Abordagem Proposta

A partir do levantamento bibliográfico e de estudos práticos tangentes à linha principal da dissertação, ou seja, as RBMs e seus modelos derivados aplicados ao domínio de vídeos e imagens, foram observados duas principais lacunas: a ausência de bancos bem delimitados para problemas de classificação de eventos de alto nível, com complexas relações entre entidades e objetos ao longo do tempo; e, a necessidade de técnicas extratoras de atributos/características capazes de capturar dependências temporais para agregar conhecimento às técnicas de *Deep Learning*.

Diante dessa observação, são propostas duas abordagens para mitigar essas lacunas. A primeira abordagem é responsável por tratar o problema da falta de bancos de dados para o domínio de eventos em vídeos. Já a segunda, visa reduzir, ou retirar, a dependência da agregação de atributos extraídos por técnicas capazes de capturar dependências temporais, como o fluxo óptico (FLEET; WEISS, 2006), por exemplo.

No que diz respeito à resolução do primeiro problema, atualmente uma área de estudo tem ganhado destaque, a transferência de aprendizado. Em linhas gerais, a TA aborda métodos para transferir o aprendizado adquirido em um problema cujos dados são abundantes, para outro com dados limitantes (pouca quantidade). É uma prática que se tornou comum para problemas no domínio de classificação que utilizam imagens (RAINA; NG; KOLLER, 2006; RAINA et al., 2007).

Já para a captura de dependências temporais, abordagens complexas baseadas em convoluções e combinação de operadores matemáticos em diferentes redes ganharam destaque, como o trabalho de Feichtenhofer et al. (FEICHTENHOFER; PINZ; ZISSERMAN, 2016) e Zhu et al. (ZHU et al., 2018). Porém, estas abordagens fazem uso de grandes e complexas redes convolucionais, requerendo grande poder computacional e tempo de treinamento.

Isto posto, a seguir são apresentados os conceitos previamente mencionados em maior profundidade, bem como a modelagem da abordagem proposta e a metodologia para este trabalho.

4.1 Adaptação de Domínio

As tarefas de classificação de ações e eventos possuem semelhanças e particularidades, apresentando desafios distintos ao treinar um modelo de aprendizado de máquina ou aprendizado profundo, como a construção de conhecimento a respeito da interação e movimentação de entidades com o passar dos *frames*. Uma abordagem interessante que pode auxiliar tarefas de reconhecimento de eventos é o aproveitamento de dados de domínios semelhantes, ou ainda a

adaptação de domínio.

A transferência de aprendizado tem recebido grande atenção nos últimos anos, principalmente com o advento do grande banco de dados ImageNet[4], e consequentemente, a adaptação de domínio, por ser uma subárea desta. Este fato possibilita o treinamento de grandes redes neurais profundas em grandes bancos dados, para posterior realização de ajuste fino destes modelos em problemas/domínios específicos.

Recentemente, novas abordagens consideraram técnicas de *Deep Learning* para a tarefa, como Tas e Koniusz (2018), que empregaram redes neurais convolucionais para reconhecimento de ações e adaptação de domínio em esqueletos corporais 3D, sob o paradigma de aprendizado supervisionado. Além disso, Liu et al. (LIU et al., 2019) propuseram um modelo para adaptação do domínio de imagens para vídeos, utilizando fusão de redes convolucionais para o reconhecimento de ações.

A adaptação de domínio estuda as possibilidades de transferência de conhecimento dos domínios de origem (domínio fonte) para diferentes contextos (domínio destino). Para ilustrar essa ideia, consideremos um carro de condução autônoma, treinado com dados de tráfego da Nova Zelândia. Ele não funcionará efetivamente nas ruas brasileiras devido às diferentes regras de sinalização e tráfego rodoviário, alto fluxo de veículos e direção do condutor à direita, por exemplo. No entanto, adaptar o conhecimento aprendido na Nova Zelândia ao Brasil pode reduzir os custos computacionais e o tempo necessário para treinar novos modelos (VENKATESWARA; CHAKRABORTY; PANCHANATHAN, 2017).

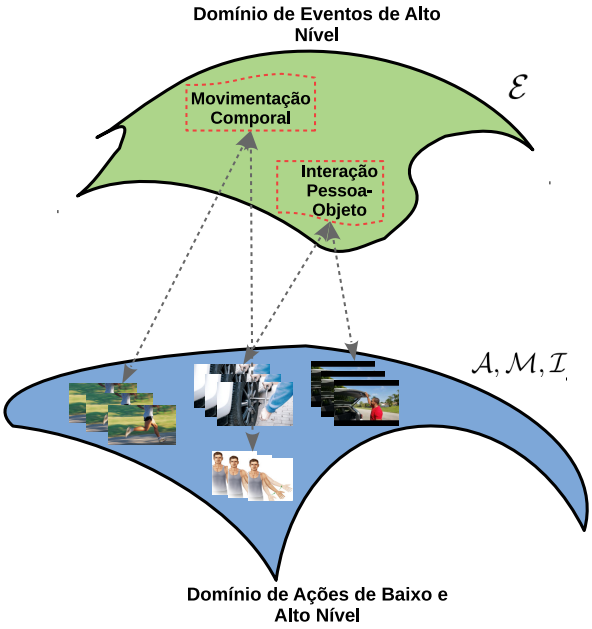
Isto posto, é possível descrever matematicamente a abordagem proposta para o problema da falta de bancos de dados no domínio de eventos em vídeos, já de forma contextualizada. A hipótese é que modelos baseados em energia são capazes de aprender atributos/características do domínio de ações que podem ser empregados para classificar eventos complexos. Portanto, a adaptação de domínio é definida como segue:

Seja Γ a tarefa de reconhecimento de eventos de alto nível, assim como $\mathcal{D}_S$ e $\mathcal{D}_T$ são os domínios de origem (*Source Domain*), o domínio de ações, e alvo (*Target Domain*), o domínio de eventos, respectivamente. Adicionalmente, o primeiro domínio é composto pelos subespaços $\mathcal{A} \in \mathbb{R}^{d_a}$, $\mathcal{M} \in \mathbb{R}^{d_m}$, e $\mathcal{I} \in \mathbb{R}^{d_i}$, onde $\{\mathcal{A}, \mathcal{M}, \mathcal{I}\} \subset \mathcal{D}_S$, enquanto o segundo é composto por $\mathcal{E} \in \mathbb{R}^{d_e}$, em que $\{\mathcal{E}\} \subset \mathcal{D}_T$. O subespaço $\mathcal{A}$ representa a base de ações com d_a dimensões, $\mathcal{M}$ representa os movimentos com d_m dimensões, $\mathcal{I}$ representa as interações entre as entidades com d_i dimensões, e finalmente, $\mathcal{E}$ é o subespaço de eventos de alto nível com d_e dimensões. A Figura 6 mostra a abordagem proposta de maneira simplificada.

Portanto, é possível definir a tarefa de adaptação de domínio (BRUZZONE; MARCON-

[4] <http://www.image-net.org/>

Figura 6 – Representação espacial dos domínios de vídeos e suas relações. $\mathcal{A}$: Ações, como correr. $\mathcal{M}$: Movimentação, por exemplo dos braços e pernas. $\mathcal{I}$: Interação, por exemplo entre pessoa e o ambiente. $\mathcal{E}$: Eventos, como a movimentação corporal composta pela movimentação de braços e pernas interagindo com o ambiente em uma corrida.



Fonte: Elaborado pelo autor.

CINI, 2009) através da Equação 22, como segue:

$$\Gamma = \{y_{\mathcal{D}_T}, f(\mathcal{D}_S)\}, \tag{22}$$

em que $y_{\mathcal{D}_T}$ representa a classe discriminativa do domínio alvo, e $f(\mathcal{D}_S)$ representa a função de mapeamento entre os domínios de origem e alvo, sendo o componente chave para que a adaptação de domínio seja eficiente e robusta.

Uma vez que as redes neurais são funções capazes de modelar relações não lineares e extrair informações intrínsecas dos dados, a proposta deste trabalho é investigar modelos baseados em energia como a função de mapeamento $f(\mathcal{D}_S)$, responsável por extrair e adaptar informações do domínio fonte $\mathcal{D}_S$ para o domínio alvo $\mathcal{D}_T$, ou seja, extrair informações relevantes de vídeos do domínio de ações para classificar eventos de alto nível.

4.2 Abordagem *Somatório*

Tratando-se de análise e processamento de vídeos para modelos de DL, dois pontos são importantes de serem destacados. O primeiro diz respeito ao processamento dos clipes, uma vez que o processamento de vídeos para modelos de DL requer a apresentação de todos

os *frames* para a técnica, tornando o processamento custoso ao considerar uma alta taxa de *frames* por clipes.

Em contrapartida, o segundo ponto diz respeito à extração de informação temporal dos vídeos, um ponto chave para a agregação de características que auxiliam na modelagem das interações entre agentes e entidades ao longo da linha do tempo nos vídeos, ou seja, com o passar dos *frames*.

Neste trabalho é apresentada uma nova abordagem (*Somatório*) aplicada aos modelos baseados em energia (Seção 3.1), com o intuito de simplificar o processamento dos *frames*, reduzir a carga computacional, e prover uma abordagem capaz de capturar informações temporais das cenas.

A abordagem *Somatório* tem como base o processamento simultâneo de todos os *frames* extraídos dos clipes, ao invés de processar *frame* a *frame* no processo de treinamento de uma RBM, por exemplo, visando reduzir a quantidade de dados apresentados para os modelos empregados e ao mesmo carregar algumas informações temporais das cenas. A intuição sobre a temporalidade vem da possibilidade de ressaltar regiões de movimentações importantes ao longo do tempo, através de um espectro de sombras que as entidades deixam ao interagir e realizar movimentos que geram essas sombras ao somar os *frames* de uma sequência.

Para proporcionar o ganho computacional previamente mencionado, a abordagem *Somatório* faz uso da agregação de todos os *frames* $\{F_1, F_2, \ldots, F_n\}$ em um único *frame* denotado por F_r, que representa a soma direta de F_1 a F_n. F_r é responsável por carregar informações espaço-temporais dos clipes, como contornos e bordas ao longo da trajetória temporal, destacadas como um espectro de movimentação na imagem resultante. Portanto, a abordagem *Somatório* para os modelos baseados em energia possui as distribuições de probabilidades condicionais da primeira camada oculta em relação a F_r como segue:

$$P(h_j = 1|\boldsymbol{F_r}) = \sigma\left(b_j + \sum_{i=1}^{m} w_{ij} F_{ri}\right), \tag{23}$$

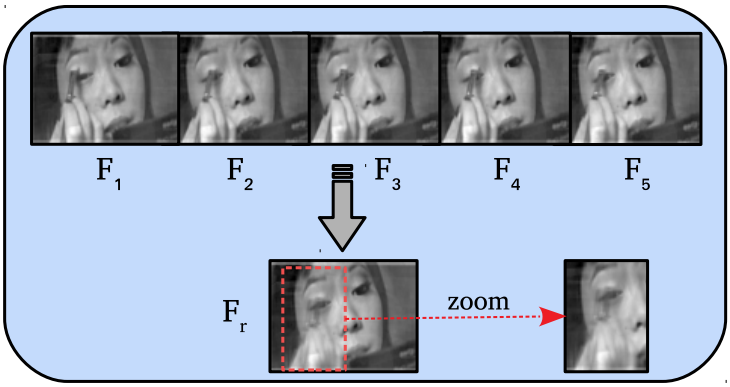
$$P(h_{1j} = 1|\boldsymbol{F_r}, \mathbf{h}_2) = \sigma\left(b_j + \sum_{i=1}^{m_1} w^1_{ij} F_{ri} + \sum_{j=1}^{n_2} w^2_{ij} h_{2j}\right). \tag{24}$$

A Equação 23 diz respeito às probabilidades condicionais dos modelos RBM e DBN, visto que a amostragem da primeira camada oculta só depende da visível, e nesse caso, F_r. já a Equação 24 representa as condicionais para uma DBM, que possui a amostragem condicionada não só à F_r, mas também à segunda camada oculta, $\mathbf{h}_2$.

A Figura 7 mostra o processo de agregação para obter F_r, assim como a região de destaque que carrega informações temporais da cena. Este processo possibilita uma completa atualização dos parâmetros das redes neurais a cada iteração, contrário ao que ocorre em uma abordagem padrão, em que os parâmetros são atualizados a cada *frame* apresentado no

treinamento. Com isso, é possível notar a diferença no que diz respeito à carga computacional, reduzindo o processamento de n *frames*, e consequentemente n atualizações em uma iteração, para apenas 1 passo.

Figura 7 – Processo de agregação de *frames* para a abordagem *Somatório*.



Fonte: Elaborado pelo autor.

Entretanto, cabem algumas observações a respeito da abordagem. A primeira, diz respeito ao possível impacto positivo quando os agentes na cena não se movimentam e/ou interagem exageradamente, destacando regiões que podem realmente ser relevantes em um curto espaço de tempo. Em contrapartida, se os agentes possuem um elevado grau de interação em um curto espaço de tempo, é possível que a soma dos *frames* acarrete em uma imagem final com muitas sombras, sem capacidade de representação eficiente das ações.

5 Experimentos

Neste capítulo é descrita a metodologia da abordagem proposta, ou seja, como os modelos baseados em energia são empregados para a tarefa de classificação de eventos de alto nível, utilizando a adaptação de domínio e os modelos *Somatório*. Além disso, são apresentados os bancos de dados utilizados para a validação da abordagem proposta, juntamente com as configurações para os experimentos.

5.1 Bancos de Dados

A partir da análise feita no Capítulo 2, Seção 2.2, acerca dos bancos de dados públicos, dois deles empregados em diversos trabalhos como *benchmark* foram escolhidos para validar a adaptação de domínio e os modelos *Somatório*. Estes são o UCF-101 e o HMDB-51, amplamente utilizados em tarefas de classificação de ações de alto nível devido às suas características desafiadoras.

Ambos bancos de dados possuem uma grande variedade de vídeos reais, extraídos do YouTube. Além de grandes variações intra-classes, possuem variações expressivas nos vídeos no que diz respeito à movimentação da câmera, aparência e movimentação dos objetos na cena, ponto de vista, fundo da cena variável e ausência de controle na iluminação.

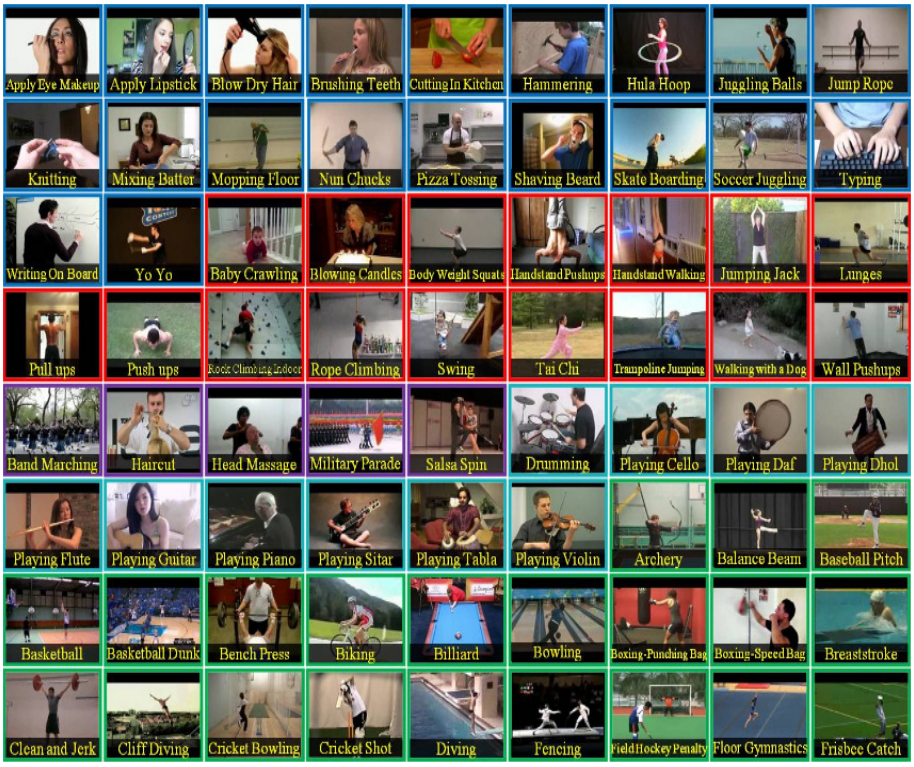
Além de serem bancos desafiadores e amplamente empregados na literatura, eles possuem outra característica intrínseca e consoante, o agrupamento das ações em eventos de alto nível. Esse fato possibilita a utilização das bases para a abordagem proposta.

Os autores do UCF-101 (SOOMRO; ZAMIR; SHAH, 2012) agruparam as 101 classes em 5 macro-categorias, que são facilmente interpretadas como os eventos de alto nível. Este mapeamento considera que as classes de ações contidas em uma macro-classe (evento) compartilham características e atributos padrões, auxiliando no reconhecimento das ações e interações. Seguindo as orientações dos autores, 5 eventos de alto nível são empregados:(0) prática de esporte(s); (1) prática musical com instrumento(s); (2) interação entre pessoa(s) e objeto(s); (3) movimentação corporal; e (4) pessoas interagindo entre si.

Na Figura 8 são apresentados *frames* de clipes aleatórios da base de dados, em que a cor da borda representa a classe do evento: verde para 0, azul claro para 1, azul para 2, vermelho para 3, e roxo para 4. Além disso, os autores fornecem os dados separados em três partições, 1, 2, e 3. Cada partição possui sua própria separação dos dados para treino e teste, sendo a partição 1 a mais utilizada em trabalhos de classificação de ações.

Para o banco de dados HMDB-51 (KUEHNE et al., 2011), os autores também agruparam as 51 classes em 5 macro-categorias, interpretadas como os eventos de alto nível. Seguindo as

Figura 8 – Exemplos de *frames* que compõe os eventos de alto nível para o banco UCF-101.



Fonte: Soomro, Zamir e Shah (2012).

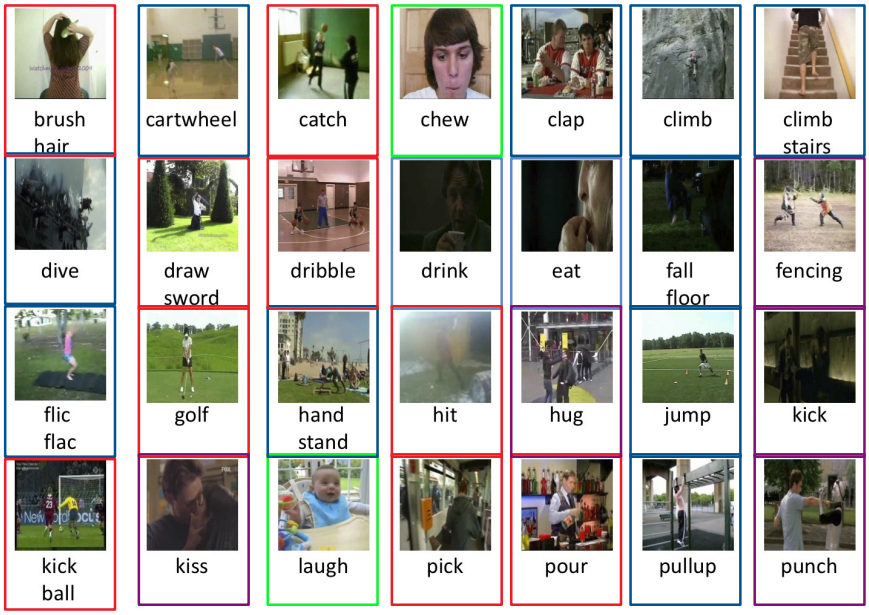
orientações dos autores, 5 eventos de alto nível são empregados:(0) expressões faciais humanas; (1) manipulação de objetos na região da face; (2) movimentação corporal; (3) interação entre pessoa(s) e objeto(s); e (4) pessoas interagindo entre si, em que os números em parêntesis representam as classes.

Adicionalmente, na Figura 9 são apresentados *frames* de clipes da base de dados, em que a cor da borda representa a classe do evento: verde para 0, azul claro para 1, azul para 2, vermelho para 3, e roxo para 4. Os autores também fornecem três partições para os dados, 1, 2, e 3, em que cada partição possui os dados separados em treino e teste.

Em ambos bancos de dados o processo de separação/aquisição dos *frames* é feito de maneira semelhante ao trabalho de Ng et al. (2015), utilizando 6 *frames* por videoclipe igualmente divididos no tempo total. Neste trabalho, os autores mostraram que 6 *frames* por vídeo são suficientes para garantir um bom desempenho, atingindo os mesmo resultados que 20 *frames*, por exemplo, além de imprimir uma menor carga computacional de processamento na tarefa de classificação de ações.

No que diz respeito ao pré-processamento, duas transformações são empregadas a partir da conversão das imagens em tons de cinza. A primeira transformação diz respeito às operações de corte, para remover regiões pretas que não carregam informações, e redimensionamento do tamanho original (240×320) para 72×96, a fim de facilitar o processamento dos modelos

Figura 9 – Exemplos de *frames* que compõe os eventos de alto nível para o banco HMDB-51.



Fonte: Kuehne et al. (2011).

baseados em energia. A segunda transformação representa a normalização dos dados usando uma distribuição Gaussiana, com média zero e variância unitária.

5.2 Configuração Experimental

No que diz respeito à configuração do hardware empregado nos experimentos, foi utilizado um Intel 2x Xeon(R) E5-2620 de 2.20GHz e 40 núcleos, com 128 GB de memória RAM, e uma placa de vídeo NVIDIA GTX 1080 Ti. Já para a configuração experimental dos modelos empregados para os respectivos bancos de dados, algumas variações foram testadas, a fim de prover maior entendimento sobre o comportamento dos modelos baseados em energia para as tarefas de classificação de eventos de alto nível utilizando a adaptação de domínio. A Tabela 2 descreve as arquiteturas utilizadas neste trabalho, bem como os hiper-parâmetros para cada modelo.

Todos os modelos destes experimentos empregaram 3 épocas de pré-treinamento no processo de extração de atributos do domínio fonte, para cada camada oculta adicionada, utilizando a abordagem de mini-lote (*mini-batch*) contendo 128 amostras de dados por lote. Já as DBMs empregam 3 épocas adicionais de treinamento a partir do pré-treinamento realizado camada a camada, como proposto por Salakhutdinov e Hinton (2012). Além disso, todos os modelos utilizam momento (*momentum*) para a atualização de seus parâmetros.

Uma vez que o intuito é utilizar os modelos baseados em energia como parte do

Modelo	**Camadas**	**Neurônios Ocultos**	**Momento**	**Taxa de Aprendizado**
RBM	1	$[2.000]$	$[0,5]$	$[1 \cdot 10^{-3}]$
DBN$_{\alpha}$	2	$[2.000 - 2.000]$	$[0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}]$
DBM$_{\alpha}$	2	$[2.000 - 2.000]$	$[0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}]$
DBN$_{\beta}$	3	$[2.000 - 2.000 - 2.000]$	$[0,5; 0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
DBM$_{\beta}$	3	$[2.000 - 2.000 - 2.000]$	$[0,5; 0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
DBN$_{\iota}$	2	$[4.000 - 4.000]$	$[0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
DBM$_{\iota}$	2	$[4.000 - 4.000]$	$[0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
DBN$_{\zeta}$	3	$[4.000 - 4.000 - 4.000]$	$[0,5; 0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
DBM$_{\zeta}$	3	$[4.000 - 4.000 - 4.000]$	$[0,5; 0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
S-RBM	1	$[2.000]$	$[0,5]$	$[1 \cdot 10^{-3}]$
S-DBN$_{\alpha}$	2	$[2.000 - 2.000]$	$[0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}]$
S-DBM$_{\alpha}$	2	$[2.000 - 2.000]$	$[0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}]$
S-DBN$_{\beta}$	3	$[2.000 - 2.000 - 2.000]$	$[0,5; 0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
S-DBM$_{\beta}$	3	$[2.000 - 2.000 - 2.000]$	$[0,5; 0,5; 0,5]$	$[1 \cdot 10^{-3}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
S-DBN$_{\iota}$	2	$[4.000 - 4.000]$	$[0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
S-DBM$_{\iota}$	2	$[4.000 - 4.000]$	$[0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
S-DBN$_{\zeta}$	3	$[4.000 - 4.000 - 4.000]$	$[0,5; 0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$
S-DBM$_{\zeta}$	3	$[4.000 - 4.000 - 4.000]$	$[0,5; 0,5; 0,5]$	$[5 \cdot 10^{-4}; 5 \cdot 10^{-4}; 5 \cdot 10^{-4}]$

Tabela 2 – Configuração experimental dos modelos empregados.

processo de classificação, cada arquitetura recebeu adicionalmente, após o pré-treinamento, duas camadas de neurônios totalmente conectados (*Fully-Connected* - FC) para possibilitar o ajuste fino sob o paradigma de aprendizado supervisionado, sendo a saída da última FC

proveniente de ativações do tipo *Softmax*, responsável por classificar os eventos de alto nível correspondentes. Uma DBM com duas camadas ocultas e duas FC é mostrada na Figura 10.

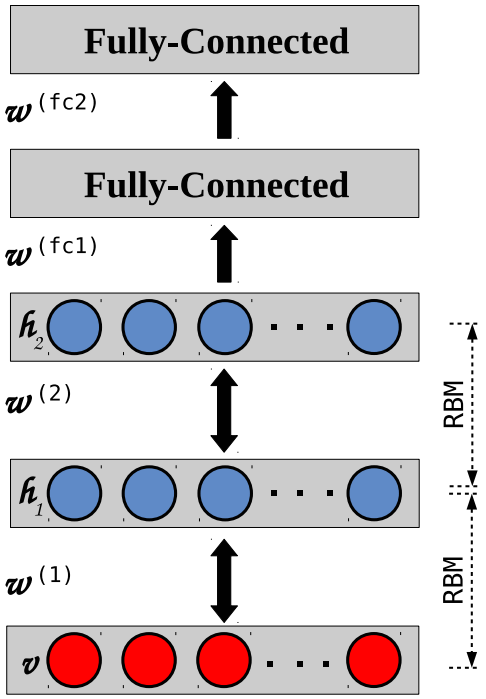


Figura 10 – DBM com duas camadas ocultas e duas FC.

O procedimento previamente descrito possibilita o aproveitamento do conhecimento dos modelos sob o paradigma de aprendizado sem supervisão, bem como a adaptação de domínio utilizando o ajuste fino com supervisão (apresentação das classes). Para isso, o novo modelo classificador foi ajustado com um otimizador bem estabelecido na literatura, o Adam (do inglês, *Adaptive momentum*) (KINGMA; BA, 2015), empregando uma taxa de aprendizado igual a $1 \cdot 10^{-3}$, por 3 épocas e o mesmo número de amostras por mini-lote (128).

As camadas FC adicionadas para o classificador podem assumir duas configurações, de acordo com a última camada oculta do modelo de energia pré-treinado, ou seja, modelos com 2.000 neurônios finais resultam em $2.000 - 1.000 - 5$ neurônios nas camadas FC, enquanto modelos com 4.000 resultam em $4.000 - 2.000 - 5$, sendo 5 o número de eventos de alto nível.

Além disso, como a tarefa de adaptação de domínio é uma subárea da transferência de aprendizado, é interessante manter o procedimento utilizado na TA para modelos pré-treinados, ou seja, "congelar" os pesos das conexões da primeira camada oculta (nesse caso, da primeira RBM do conjunto) e realizar um ajuste fino sutil das camadas subsequentes, com uma baixa taxa de aprendizado igual a $1 \cdot 10^{-6}$.

Por fim, todo o procedimento de ajuste fino dos modelos é guiado por uma função de custo, e neste caso, a entropia-cruzada (*cross-entropy*) foi empregada por ser amplamente utilizada em problemas de classificação, levando em consideração o erro de classificação (BISHOP, 1995). A métrica de avaliação final é a acurácia, e os modelos foram executados 6 vezes a fim

de mitigar a natureza estocástica das técnicas.

É importante salientar que mais épocas de pré-treinamento e ajuste fino foram testadas em estudos preliminares, porém, as diferenças no desempenho preditivo dos modelos não foram significativas, ao passo que o tempo de execução foi impactado negativamente, provavelmente pelas limitações inerentes às técnicas empregadas. Portanto, apenas os modelos com 3 épocas foram mantidos para os experimentos e análises finais, os quais representam os melhores cenários de "custo-benefício".

6 Resultados Experimentais

Este capítulo é dedicado à apresentação dos resultados experimentais abordando a resolução dos problemas citados, utilizando a adaptação de domínio e a proposta de processamento de *frames* de vídeos nos modelos baseados em energia para os bancos UCF-101 e HMDB-51.

O capítulo é separado em duas seções, a primeira aborda a extração de atributos utilizando o aprendizado não supervisionado, analisando a empregabilidade destes para a tarefa de adaptação de domínio e posterior classificação de eventos complexos. Já a segunda seção diz repeito aos resultados de classificação, considerando o processo de ajuste fino dos modelos *Somatório* e os empregados como linha de base.

6.1 Aprendizado Não Supervisionado

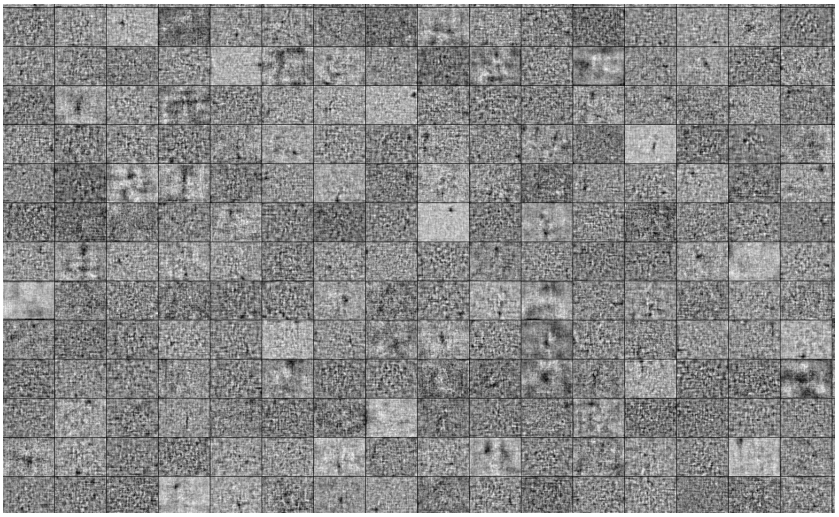
A análise visual dos pesos aprendidos por uma técnica de DL, sob o paradigma não supervisionado, é frequentemente uma boa maneira de analisar o processo de aprendizado, uma vez que é possível observar como as informações de baixo e alto nível são capturadas, se aproximando do processamento no córtex visual humano.

Frente aos modelos empregados nestes estudos preliminares, é possível analisar as conexões das primeiras camadas de duas principais arquiteturas, diferindo na quantidade de neurônios da primeira camada oculta. Portanto, a Figura 11 mostra a visualização referente às conexões da primeira camada das variantes S-DBN (a) e DBN (b), com 2.000 neurônios ocultos cada. Optou-se por mostrar apenas os pesos aprendidos por modelos referentes ao banco de dados UCF-101, uma vez que é o mais promissor para a tarefa devido à quantidade de videoclipes ser maior que os encontrados no HMDB-51.

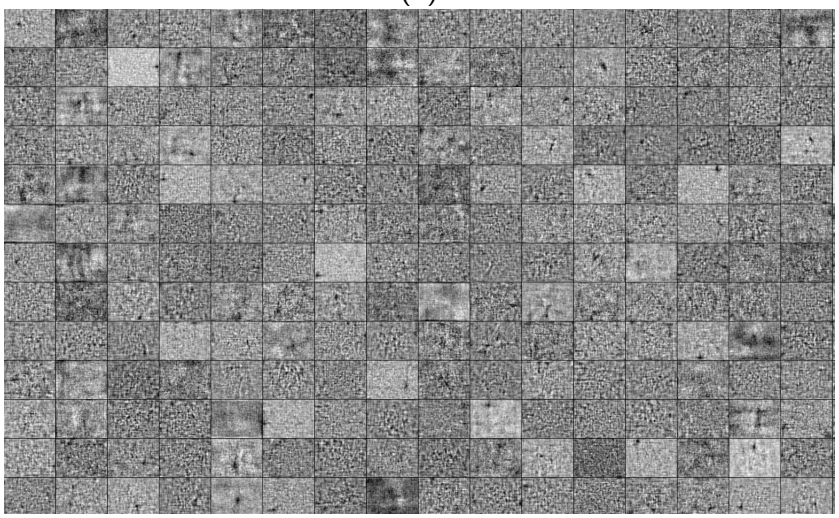
A partir da Figura 11, pode-se observar que o modelo *Somatório* foi capaz de aprender atributos mais informativos e menos saturados, isto é, menos regiões majoritariamente pretas ou brancas que representam valores muito grandes ou pequenos para as sinapses. Além disso, o modelo S-DBN mostrou maior capacidade em aprender conexões que favorecem a combinação linear destas para a formação de conhecimento de alto nível em camadas subsequentes, representados por pequenos contornos e pontos dispersos ao longo dos pesos.

O comportamento descrito anteriormente pode influenciar positivamente na adaptação de domínio, resultando em um bom desempenho preditivo dos eventos complexos uma vez que os neurônios ocultos estão carregando mais informações, evitando possíveis saturações em suas ativações, fato que aumenta o poder de generalização das redes.

Adicionalmente, a Figura 12 mostra algumas conexões da primeira camada oculta dos modelos (a) S-DBN e (b) DBN com 4.000 neurônios. É possível notar que o modelo *Somatório*



(a)



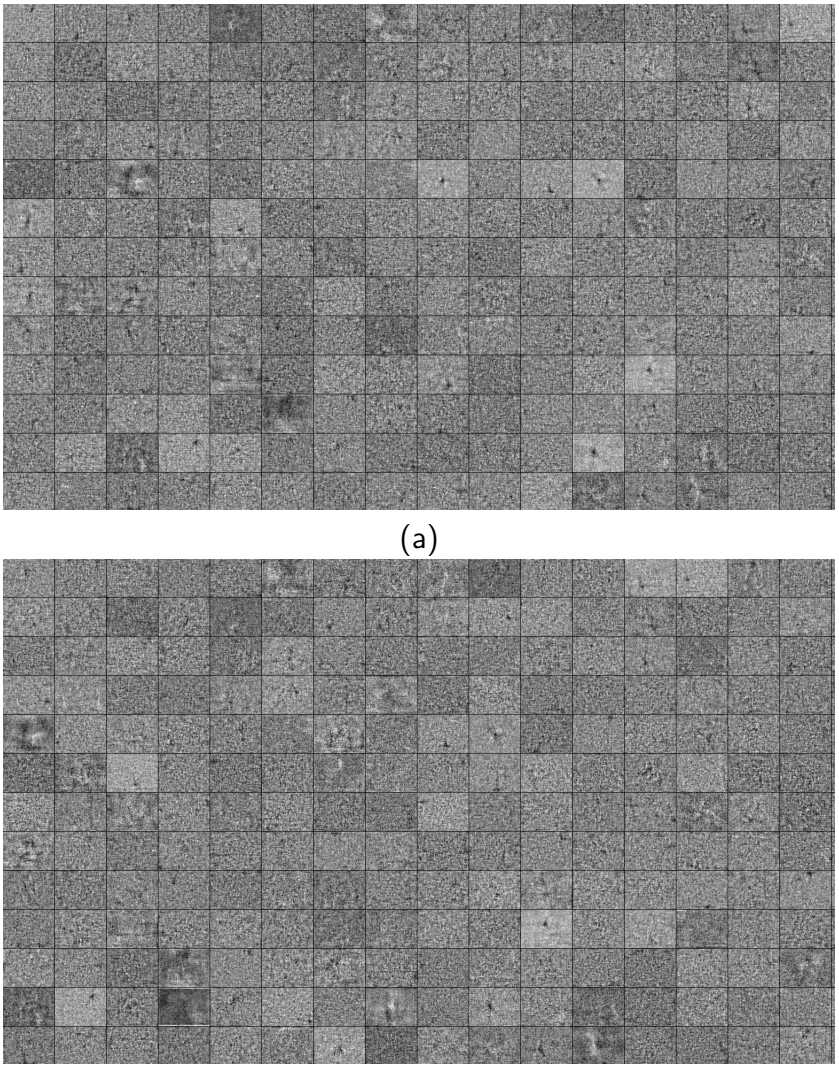
(b)

Figura 11 – Primeira camada de pesos de DBNs considerando os modelos (a) S-DBN e (b) DBN com 2.000 neurônios ocultos.

tende a manter o comportamento discutido anteriormente, ou seja, modelar sinapses com menor propensão à saturação das ativações que o modelo padrão.

As análises anteriores também foram feitas para as DBMs, padrão e *Somatório*, como mostrado nas Figuras 13 e 14 em que são representadas algumas conexões dentre os 2.000 e 4.000 neurônios ocultos, respectivamente. A partir da Figura 13, pode-se observar que o modelo S-DBM possui conexões menos saturadas que a DBM, indicando certa esparsidade ao modelo *Somatório*, podendo conferir maior capacidade de generalização a este. Regiões esparsas são representadas por tonalidades mais claras e/ou próximas da cor branca.

Por fim, a partir da Figura 14, observar-se novamente que o modelo S-DBM possui conexões menos saturadas que a DBM. Além disso, o modelo *Somatório* parece possuir menos regiões com valores aleatórios, ou seja, possuem coloração suavizada e menos caótica em alguns quadros. De maneira geral, ambos os modelos com 4.000 neurônios indicam maior



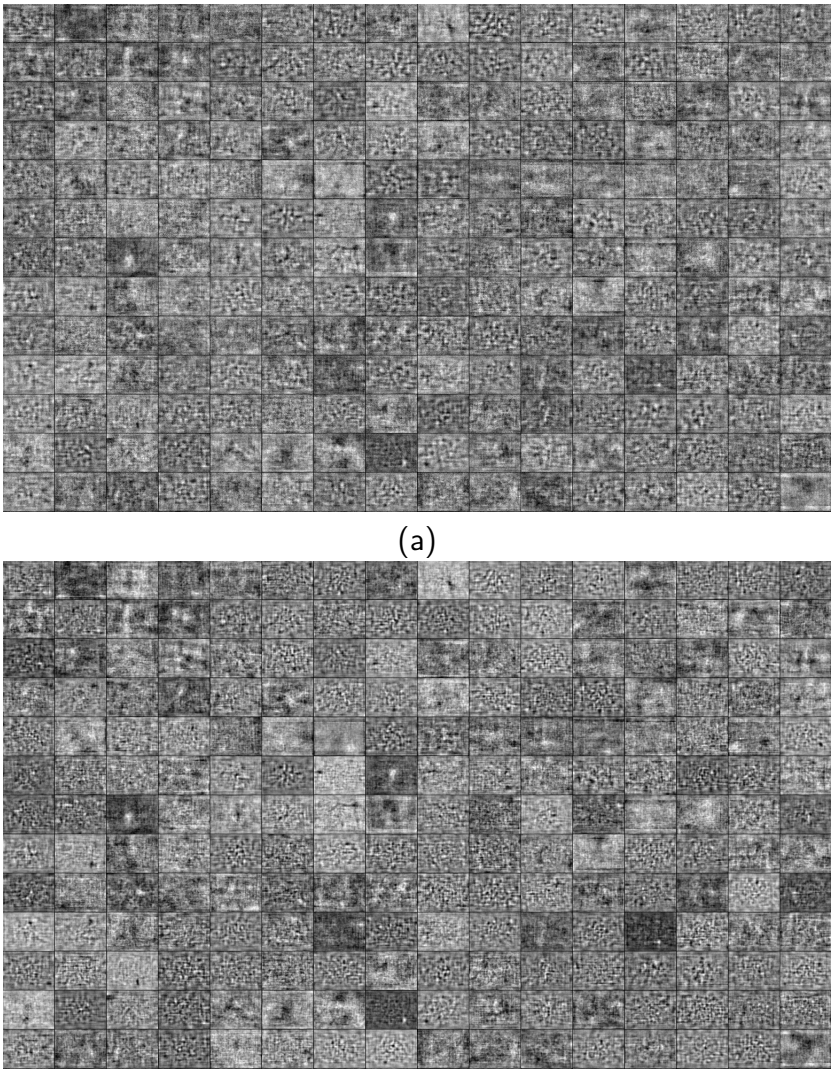
(a) (b)

Figura 12 – Primeira camada de pesos de DBNs considerando os modelos (a) S-DBN e (b) DBN com 4.000 neurônios ocultos.

suscetibilidade à agregação de informação do domínio de ações, fato que pode favorecer a adaptação de domínio para classificação de eventos de alto nível.

6.2 Avaliação dos Modelos para Classificação de Eventos de Alto Nível

Nesta seção são apresentados os resultados em duas subseções, cada uma respectiva a um banco de dados empregado, ou seja, UCF-101 e HMDB-51. Além disso, algumas observações importantes são tratadas acerca do impacto computacional de cada modelo utilizado neste trabalho.



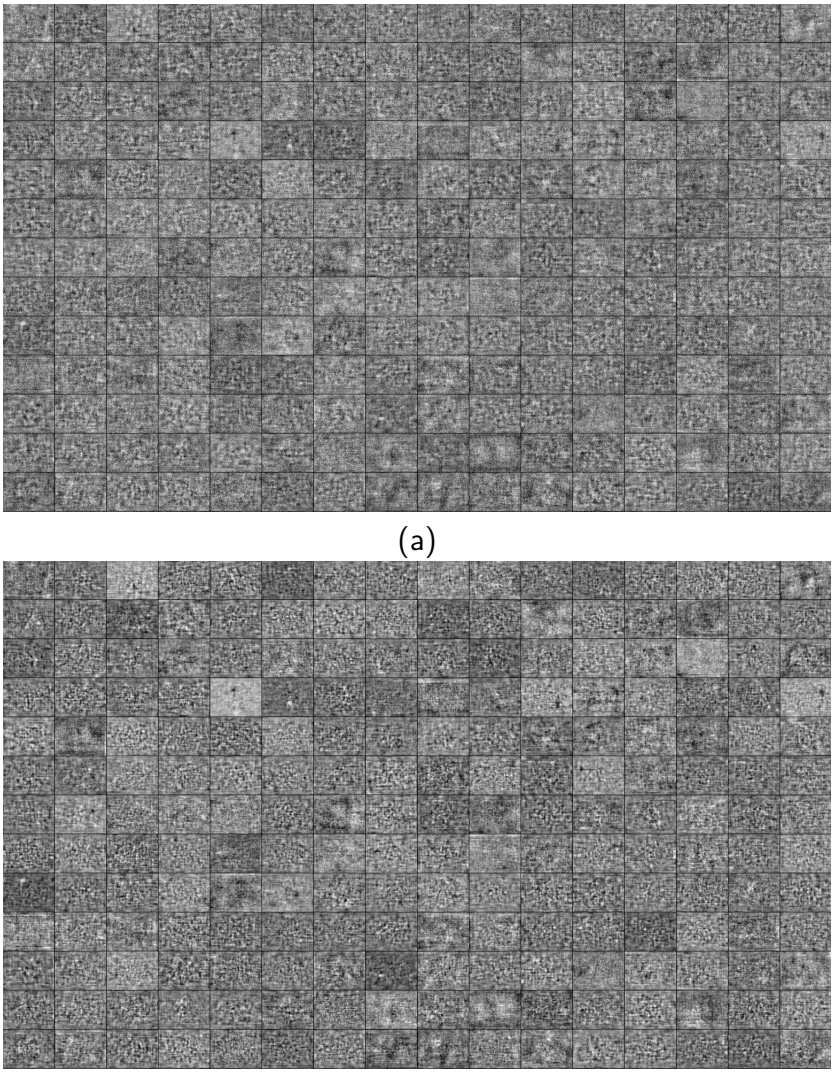
(a) (b)

Figura 13 – Primeira camada de pesos de DBMs considerando os modelos (a) S-DBM e (b) DBM com 2.000 neurônios ocultos.

6.2.1 Banco de Dados UCF-101

Considerando a tarefa de classificação de eventos de alto nível a partir de ações em vídeos, a Tabela 3 mostra o desempenho preditivo para todos os modelos elucidados na Seção 5.2, executados 6 vezes cada. O valor destacado em negrito indica a maior acurácia média atingida entre todos os modelos empregados. Além disso, é apresentado o tempo de execução considerando as 3 épocas de pré-treinamento em todas as camadas ocultas presentes nos modelos, a fim de analisar o impacto computacional e a eficiência de cada modelo testado.

A partir da Tabela 3, é possível observar resultados promissores, principalmente para a abordagem *Somatório* dos modelos RBM e DBN. Analisando o modelo S-RBM, vê-se que este atingiu maiores taxas de acerto média que sua versão padrão (RBM), com valores de $42,48\%$ contra $38,71\%$, ou seja, $3,77$ pontos de diferença na acurácia. Adicionalmente, a S-RBM tem a menor carga computacional, aproximadamente 14% menos tempo de execução. Destacando



(a)

(b)

Figura 14 – Primeira camada de pesos de DBMs considerando os modelos (a) S-DBM e (b) DBM com 4.000 neurônios ocultos.

que daqui em diante, a diferença percentual entre as acurácias dos modelos representa o valor absoluto médio da versão *Somatório* subtraído da sua versão padrão, enquanto para o tempo de execução o valor percentual representa o quociente entre o tempo médio da abordagem *Somatório* e o modelo padrão.

Considerando a segunda comparação entre arquiteturas, ou seja, os modelos α, a versão *Somatório* (S-DBN$_\alpha$) superou os resultados de acurácia média da DBN em praticamente 7%, e de todos modelos desta linha de base, indicando um aumento representativo. Entretanto, o modelo padrão (DBN) não conseguiu superar os resultados da sua versão mais simples (RBM) em termos de acurácia, ao passo que a DBM$_\alpha$ superou as versões padrões anteriores. Além disso, o tempo de execução para a S-DBN$_\alpha$ foi 30% menor que a DBN$_\alpha$, impactando positivamente em um treinamento menos custoso. Em contrapartida, os tempos de execução das DBMs foram maiores, fato esperado devido ao processo de pre-treinamento requerido para

Modelo	**Acurácia**	**Tempo (min)**
RBM	$38,71 \pm 1,04$	$315,00 \pm 5,00$
S-RBM	$42,48 \pm 0,94$	$270,00 \pm 5,00$
DBN$_{\alpha}$	$37,72 \pm 4,67$	$765,00 \pm 5,00$
S-DBN$_{\alpha}$	$44,66 \pm 1,28$	$540,00 \pm 5,00$
DBM$_{\alpha}$	$43,40 \pm 4,25$	$1.260,00 \pm 5,00$
S-DBM$_{\alpha}$	$44,55 \pm 4,38$	$961,00 \pm 5,00$
DBN$_{\beta}$	$40,55 \pm 3,54$	$1.215,00 \pm 5,00$
S-DBN$_{\beta}$	$44,80 \pm 2,02$	$810,00 \pm 5,00$
DBM$_{\beta}$	$41,59 \pm 4,23$	$1.785,00 \pm 5,00$
S-DBM$_{\beta}$	$41,27 \pm 4,26$	$1.295,00 \pm 5,00$
DBN$_{\iota}$	$41.92 \pm 2,65$	$775,00 \pm 6,00$
S-DBN$_{\iota}$	$\mathbf{45,01 \pm 1,39}$	$550,00 \pm 6,00$
DBM$_{\iota}$	$41,64 \pm 4,30$	$2.110,00 \pm 6,00$
S-DBM$_{\iota}$	$43,17 \pm 4,38$	$1685,00 \pm 6,00$
DBN$_{\zeta}$	$42,33 \pm 4,51$	$1.225,00 \pm 6,00$
S-DBN$_{\zeta}$	$44,87 \pm 2,81$	$820,00 \pm 6,00$
DBM$_{\zeta}$	$41,61 \pm 4,27$	$2.563,00 \pm 6,00$
S-DBM$_{\zeta}$	$44,89 \pm 4,36$	$1.957,00 \pm 6,00$

Tabela 3 – Acurácias médias (%) e tempo de execução (minutos) para o banco UCF-101.

a inicialização da rede.

A respeito dos modelos β, o mesmo comportamento pode ser observado, ou seja, a versão *Somatório* superando os resultados da DBN padrão em aproximadamente 4% de acurácia

média, e 33% menos tempo de execução. Mesmo com bons resultados, a adição de outra camada oculta ao modelo *Somatório* não gerou um aumento de desempenho muito expressivo, uma vez que a S-DBN$_\beta$ teve resultados próximos à S-DBN$_\alpha$. Já as DBMs β (*Somatório* e padrão) tiveram resultados piores que suas respectivas versões α, entretanto, mesmo com as observações anteriores, os modelos *Somatório* tiveram menor tempo de execução que suas versões padrão.

A respeito da quarta arquitetura (modelos ι), a S-DBN$_\iota$ superou sua versão padrão, atingindo a notável acurácia média de $45,01\%$, o maior valor médio frente a todos os modelos do estudo. Além disso, o tempo de execução para a S-DBN$_\iota$ foi 30% melhor que sua versão padrão. Aqui, os modelos *Somatório* mostraram que a adição de mais neurônios ocultos pode ser benéfico para a melhoria de desempenho geral na tarefa de adaptação de domínio e classificação de eventos de alto nível, quando comparados às versões padrão (DBN e DBM). Os resultados para a DBM$_\iota$ não foram tão satisfatórios quanto se esperava, principalmente ao comparar essa versão com a de 2.000 neurônios ocultos (α), a qual atingiu melhor resultado de acerto médio e melhor tempo de execução.

Por fim, os modelos ζ alcançaram praticamente o mesmo resultado dos modelos ι, principalmente para a S-DBN, que atingiu a acurácia média de $44,87\%$. Este resultado indica que a adição de mais neurônios e mais camadas ocultas pode realmente impactar positivamente o desempenho destes modelos. Porém, com o aumento da complexidade das arquiteturas estas acabam consumindo um tempo de pré-treinamento maior, podendo inviabilizar o ganho de desempenho. Novamente, os tempos de execução dos modelos *Somatório* foram melhores que as versões padrão. Para estes últimos experimentos a S-DBM se sobressaiu sobre os demais, atingindo a marca de $44,89\%$ de acerto médio, a maior para as DBMs até então, possivelmente por conta da maior capacidade de extração de conhecimento conferida pelo aumento de neurônios e camadas ocultas.

Não obstante, é essencial destacar que os modelos S-DBN não tiveram dificuldade em atingir melhores acurácias médias que suas versões padrão e as RBMs, fato que está ligado diretamente à capacidade de abstração hierárquica de camadas subjacentes. Os modelos *Somatório* conseguiram melhorar o *lower bound* (limite inferior) da função *Log-likelihood*, resultando em conexões úteis para a tarefa discriminativa feita após o ajuste fino. Já para as DBMs, na maioria das arquiteturas a versão *Somatório* conferiu melhoria de desempenho frente à versão padrão, porém, a melhoria não foi tão expressiva quanto nas DBNs.

6.2.2 Banco de Dados HMDB-51

Considerando a tarefa de classificação de eventos de alto nível a partir de ações em vídeos, a Tabela 4 mostra o desempenho preditivo para todos os modelos elucidados na Seção 5.2, executados 6 vezes cada. O valor destacado em negrito indica a maior acurácia média atingida entre todos os modelos empregados. Além disso, é apresentado o tempo de

execução considerando as 3 épocas de pré-treinamento em todas as camadas ocultas presentes nos modelos, a fim de analisar o impacto computacional e a eficiência de cada modelo testado.

Modelo	**Acurácia**	**Tempo (min)**
RBM	$34,60 \pm 3,90$	$45,00 \pm 5,00$
S-RBM	$35,19 \pm 4,24$	$30,00 \pm 5,00$
DBN$_{\alpha}$	$34,49 \pm 3,98$	$144,00 \pm 5,00$
S-DBN$_{\alpha}$	$37,68 \pm 4,19$	$132,00 \pm 5,00$
DBM$_{\alpha}$	$33,78 \pm 4,08$	$234,00 \pm 5,00$
S-DBM$_{\alpha}$	$36,53 \pm 4,23$	$162,00 \pm 5,00$
DBN$_{\beta}$	$33,83 \pm 4,06$	$216,00 \pm 5,00$
S-DBN$_{\beta}$	$38,70 \pm 4,33$	$198,00 \pm 5,00$
DBM$_{\beta}$	$\mathbf{39,04 \pm 4,31}$	$336,00 \pm 5,00$
S-DBM$_{\beta}$	$37,72 \pm 4,26$	$238,00 \pm 5,00$
DBN$_{\iota}$	$34,41 \pm 4,03$	$150,00 \pm 6,00$
S-DBN$_{\iota}$	$38,23 \pm 4,25$	$138,00 \pm 6,00$
DBM$_{\iota}$	$38,42 \pm 4,22$	$246,00 \pm 6,00$
S-DBM$_{\iota}$	$37,96 \pm 4,19$	$170,00 \pm 6,00$
DBN$_{\zeta}$	$34,53 \pm 4,04$	$225,00 \pm 6,00$
S-DBN$_{\zeta}$	$38,86 \pm 4,26$	$207,00 \pm 6,00$
DBM$_{\zeta}$	$36,35 \pm 4,20$	$351,00 \pm 6,00$
S-DBM$_{\zeta}$	$38,02 \pm 4,30$	$249,00 \pm 6,00$

Tabela 4 – Acurácias médias (%) e tempo de execução (minutos) para o banco HMDB-51.

A partir da Tabela 4, é possível observar resultados interessantes e destoantes ao

considerar as análises da seção anterior. A abordagem *Somatório* para as RBMs atingiu uma taxa de acerto média ligeiramente maior que sua versão padrão (RBM), com valores de $35,19\%$ contra $34,60\%$. Além disso, nota-se rapidamente que os tempos de execução de todos os modelos são muito menores que os obtidos na Tabela 3, por conta da quantidade de videoclipes do banco HMDB-51. A S-RBM tem a menor carga computacional, aproximadamente 33% menor que a RBM.

A respeito da segunda comparação entre arquiteturas, ou seja, os modelos α, a versão *Somatório* superou os resultados de acurácia média da DBN em praticamente 3%, indicando um aumento representativo. Porém, o modelo padrão não conseguiu superar os resultados da sua versão mais simples (RBM) em termos de acurácia. Além disso, o tempo de execução para a S-DBN$_\alpha$ foi aproximadamente 8% menor que a DBN$_\alpha$, impactando positivamente em um treinamento menos custoso. As DBMs também foram impactadas positivamente pela abordagem *Somatório*, tanto em acurácia quanto em tempo de execução, porém, seus acertos médios foram menores que os alcançados pelas DBNs para ambas arquiteturas.

A respeito dos modelos β, o mesmo comportamento pode ser observado para as DBNs, ou seja, a versão *Somatório* superando os resultados da DBN padrão em aproximadamente $4,90\%$ de acurácia média, e $8,33\%$ menos tempo de execução. Além disso, a adição de outra camada oculta ao modelo *Somatório* gerou um ligeiro aumento de desempenho, aproximadamente 1% maior que a S-DBN$_\alpha$. Adicionalmente, observa-se uma quebra de padrão com as DBMs, ou seja, o modelo padrão se sobressaiu frente à S-DBM, atingindo a taxa de acerto média de $39,04$ contra $37,72\%$, a maior acurácia média de todos os experimentos com o banco HMDB-51.

A respeito da quarta arquitetura (modelos ι), nota-se um comportamento muito semelhante à análise feita anteriormente, ou seja, o modelo S-DBN$_\iota$ superou sua versão padrão, atingindo a acurácia média de $38,23\%$ contra $34,41\%$, ao passo que a DBM$_\iota$ superou sua versão *Somatório* em aproximadamente $0,5\%$. Porém, o tempo de execução para a S-DBM$_\iota$ foi 30% melhor que sua versão padrão.

Por fim, os modelos ζ alcançaram praticamente o mesmo resultado dos modelos ι, principalmente para as DBNs *Somatório* e padrão. Em contrapartida, as DBMs tiveram uma ligeira queda de desempenho na taxa de acerto média em relação aos modelos ι, que possuem uma camada oculta a menos, enquanto a S-DBM praticamente manteve seu desempenho preditivo. Estes resultados indicam que a adição de mais neurônios e camadas ocultas pode não ser muito benéfico para o desempenho dos modelos quando não há uma grande quantidade de dados. O comportamento do tempo de execução manteve-se o mesmo, como nas arquiteturas anteriores.

7 Conclusão

A inteligência artificial tornou-se uma ferramenta importante no dia a dia de todos, apoiada principalmente pela elevada conectividade dos dispositivos tecnológicos mais recentes. Diante disso, diversas aplicações têm ganhado destaque e são foco de diversos estudos, dentre elas a classificação de ações e eventos em vídeos, para os mais variados setores os quais o homem está inserido, direta ou indiretamente.

Neste trabalho investigou-se o problema da classificação de eventos de alto nível em vídeos, utilizando redes neurais baseadas em energia, como *Restricted Boltzmann Machines*, *Deep Belief Networks*, e *Deep Boltzmann Machines*. A hipótese estabelecida é que estes modelos são capazes de aprender e extrair atributos/características do domínio de ações que podem ser empregados para classificar eventos complexos, ou de alto nível, utilizando a adaptação de domínio em conjunto aos paradigmas de aprendizado não supervisionado e supervisionado.

Para a aplicação em questão, foi proposta a metodologia *Somatório* que visa simplificar o processamento dos vídeos e conferir robustez aos modelos diante da variação espaço-temporal que ocorre naturalmente em vídeos reais, como nos bancos de dados investigados. Os resultados experimentais mostraram que os modelos baseados em energia foram capazes de cumprir a tarefa de classificação de eventos de alto nível utilizando a adaptação de domínio com a metodologia proposta. Deste modo, pode-se confirmar a hipótese da dissertação, ou seja, os modelos empregados foram capazes de reconhecer e classificar eventos de alto nível em bancos de dados de vídeos. Além disso, alguns pontos positivos (+) e negativos (-) são importantes de serem destacados:

(+) Os modelos utilizados para o problema são relativamente simples;

(+) Utilizar as redes neurais como função de mapeamento entre domínios foi satisfatório;

(+) A área de estudo de eventos em vídeos teve a introdução de redes baseadas em energia de maneira extensiva;

(+) Duas bibliotecas em Python foram geradas junto do trabalho final, uma específica para bancos de dados de imagens (*learnergy*), e uma para bancos de dados de vídeos (*learnergy4video*), disponíveis nos respectivos links: <https://github.com/gugarosa/learnergy> e <https://github.com/MateusRoder/learnergy4video>.

(-) O banco de dados HMDB-51 possui poucos videoclipes, o que pode ter dificultado o aprendizado das redes neurais;

(-) O custo computacional foi elevado, mesmo com a abordagem *Somatório*;

(-) Os modelos utilizados possuem arquiteturas simples, dificultando ganhos muito expressivos de desempenho;

7.1 Trabalhos Futuros

Diante do conteúdo desenvolvido, diversas dificuldades foram encontradas, como o treinamento de modelos que não utilizam operadores de convolução para tratar vídeos, e a extração de informação contida no espaço-tempo proveniente de uma sequência de *frames* extraídos de videoclipes. Também é possível citar o grande tempo de treinamento quando o banco de dados é grande, mesmo utilizando o processamento em GPUs e mini-lotes relativamente grandes. Entretanto, de forma positiva, algumas oportunidades foram encontradas, possibilitando o levantamento de pontos de interesse para trabalhos futuros:

- Implementação de operações de convolução para as redes baseadas em energia;
- Investigar técnicas de regularização nas redes baseadas em energia para o domínio explorado;
- Explorar técnicas de aumento artificial de dados (*Data Augmentation*) para classes minoritárias e/ou bancos pequenos;
- Utilizar ambos bancos de dados para realizar o pré-treinamento e apenas um para o ajuste fino, representando o paradigma de aprendizado fracamente supervisionado.

8 Trabalhos Desenvolvidos

No decorrer dos estudos dos tópicos relacionados à dissertação, e da implementação dos algoritmos, surgiram possibilidades de expandir e propor algumas abordagens que culminaram em trabalhos importantes desenvolvidos paralelamente, sendo estes elencados nas próximas seções.

8.1 *Learnergy: Energy-based Machine Learners*

Ao longo dos últimos anos, as técnicas de aprendizado de máquina foram amplamente incentivadas no contexto de arquiteturas de aprendizado profundo. Um importante algoritmo denominado Máquinas de Boltzmann Restritas (um modelo de rede neural artificial) emprega conceitos de natureza baseada em energia e probabilística para lidar com as mais diversas aplicações, como classificação, reconstrução e geração de imagens e sinais. No entanto, observa-se que essas redes não são adequadamente renomadas em comparação com outras técnicas de aprendizado profundo bem conhecidas, como por exemplo, Redes Neurais Convolucionais.

Este comportamento promove certa escassez de pesquisas e principalmente implementações eficientes na literatura, dificultando a compreensão suficiente desses sistemas baseados em energia. Portanto, neste artigo, propomos um *framework* em linguagem Python para o contexto de arquiteturas baseadas em energia, denominado *Learnergy*. Essencialmente, *Learnergy* é construído utilizando o PyTorch para fornecer um ambiente mais amigável e um espaço de trabalho de prototipagem mais rápido e, possivelmente, além de possibilitar o uso de GPUs para cálculos computacionais, acelerando seu tempo de execução.

Referência do trabalho:

RODER, M.; de ROSA, G. H.; PAPA, J. P. Learnergy: Energy-based Machine Learners. *arXiv preprint arXiv:2003.07443*, 2020.

8.2 *Intestinal Parasites Classification Using Deep Belief Networks*

Atualmente, aproximadamente 4 bilhões de pessoas estão infectadas por parasitas intestinais em todo o mundo. As doenças causadas por estes constituem um problema de saúde pública na maioria dos países tropicais, levando a distúrbios físicos e mentais e até a morte de crianças e indivíduos imunodeficientes.

Embora sujeita a altas taxas de erro, a inspeção visual humana ainda é responsável pela grande maioria dos diagnósticos clínicos. Nos últimos anos, alguns trabalhos abordaram a

classificação inteligente de parasitas intestinais auxiliados por computador, mas eles geralmente sofrem de classificação incorreta devido a semelhanças entre parasitas e impurezas fecais.

Neste artigo, *Deep Belief Networks* foram aplicadas no contexto da classificação automática de parasitas intestinais. Experimentos realizados em três conjuntos de dados compostos por ovos, larvas e protozoários forneceram resultados promissores, mesmo considerando classes desequilibradas e também impurezas fecais.

Além da aplicação de DBNs e RBMs para o problema de classificação, uma nova abordagem para mitigar o desbalanceamento de classes foi proposta utilizando RBMs. Esta abordagem consiste em treinar RBMs específicas para as classes com menos exemplos e utilizar o aspecto generativo das redes para gerar imagens sintéticas que possam ser utilizadas para compor o novo banco de dados, assemelhando-se da abordagem de aumento artficial de dados (*Artificial Data Augmentation*), porém com características estocásticas que conferem variação aos dados de entrada e possibilita a utilização destes para treinar um modelo geral.

Referência do trabalho:

RODER, M.; PASSOS JUNIOR, L. A.; RIBEIRO, L. C. F.; BENATO, B. C.; FALCÃO, A. X.; PAPA, J. P. Intestinal Parasites Classification Using Deep Belief Networks. *In: 19th International Conference on Artificial Intelligence and Soft Computing*, Zakopane. ICAISC 2020: Artificial Intelligence and Soft Computing. Springer, v. 12415. p. 242-251, doi: 10.1007/978-3-030-61401-0_23, 2020.

8.3 *A Layer-Wise Information Reinforcement Approach to Improve Learning in Deep Belief Networks*

Com o advento da *Deep Learning*, o número de trabalhos propondo novos métodos ou aprimorando os existentes aumentou exponencialmente nos últimos anos. Nesse cenário, modelos " muito profundos" emergiram, esperando que extraíssem atributos mais significativos e abstratos ao mesmo tempo, possibilitando um melhor desempenho. No entanto, esses modelos sofrem com o problema de desaparecimento do gradiente (*Gradient Vanishing*), ou seja, os valores de retropropagação se tornam muito próximos de zero nas camadas mais superficiais das redes, fazendo com que o aprendizado fique estagnado.

O problema mencionado foi superado no contexto de redes neurais convolucionais, criando "conexões de atalho" entre blocos de camadas, em uma estrutura de aprendizado residual profunda. No entanto, uma técnica muito popular de aprendizado profundo chamada *Deep Belief Networks* ainda sofre com o desaparecimento do gradiente ao lidar com tarefas discriminativas. Portanto, este trabalho propôs a *Residual Deep Belief Network*, que considera o reforço de informações camada por camada para melhorar a extração de atributos e a retenção de conhecimento, que suportam melhor desempenho discriminativo. Experimentos realizados em

três conjuntos de dados públicos demonstram sua robustez em relação à tarefa de classificação de imagens binárias.

Referência do trabalho:

RODER, M.; PASSOS JUNIOR, L. A.; RIBEIRO, L. C. F.; PEREIRA, C. R.; PAPA, J. P. A Layer-Wise Information Reinforcement Approach to Improve Learning in Deep Belief Networks. *In: 19th International Conference on Artificial Intelligence and Soft Computing*, Zakopane. ICAISC 2020: Artificial Intelligence and Soft Computing. Cham: Springer, v. 12415. p. 231-241, doi: 10.1007/978-3-030-61401-0_22, 2020.

8.4 *Fine-Tuning Temperatures in Restricted Boltzmann Machines Using Meta-Heuristic Optimization*

As *Restricted Boltzmann Machines* (RBMs) são redes neurais estocásticas usadas principalmente para reconstrução de imagens e aprendizado/extração de atributos sem supervisão. Uma versão aprimorada, a RBM baseada em temperatura (T-RBM), considera um novo parâmetro de temperatura durante o processo de aprendizado que influencia a ativação dos neurônios.

No entanto, a principal vulnerabilidade de tais modelos diz respeito à seleção adequada da temperatura, o que pode levá-los a treinamento inadequado ou à adaptação excessiva quando a temperatura é configurada incorretamente, limitando a rede de prever ou trabalhar de maneira eficaz com dados não vistos.

Este artigo aborda o problema de selecionar adequadamente a temperatura de um sistema através de um processo de otimização meta-heurística, levando em consideração o erro quadrático médio. Técnicas meta-heurísticas, como otimização por enxame de partículas (*Particle Swarm Optimization*), algoritmo de morcegos (*Bat Algorithm*) e colônia artificial de abelhas (*Artificial Bee Colony*), foram empregadas para encontrar valores adequados para o parâmetro de temperatura.

Além disso, para fins de comparação, três valores de temperatura padrão e uma busca aleatória foram empregados como base comparativa. Os resultados mostraram que a otimização da T-RBM com as meta-heurísticas foi adequada, principalmente devido à complexidade da função de aptidão (dependente das conexões da RBM), o que torna o ajuste fino das temperaturas uma tarefa não trivial.

Referência do trabalho:

RODER, M.; de ROSA, G. H.; PAPA, J. P.; BREVE, F. A. Fine-Tuning Temperatures in Restricted Boltzmann Machines Using Meta-Heuristic Optimization. *In: 2020 IEEE Congress on Evolutionary Computation (CEC)*, Glasgow. 2020 IEEE Congress on Evolutionary Computation

(CEC), p. 1-8, doi: 10.1109/cec48606.2020.9185505, 2020.

8.5 *Harnessing Particle Swarm Optimization Through Relativistic Velocity*

No século passado, as percepções de Albert Einstein sobre o mundo provocaram uma revolução na compreensão do universo. Em sua teoria da relatividade geral, ele descreve o contínuo espaço-tempo, um conceito capaz de explicar vários fenômenos, que vão da gravidade a buracos negros e supernovas. Além disso, também fornece um conjunto de formulações para generalizar conceitos de física clássica para acomodar as noções relativísticas.

Enquanto isso, vários matemáticos têm trabalhado em ferramentas de otimização com o objetivo de resolver problemas complexos associados a um grande número de variáveis. Atualmente, apesar do poder computacional, muitas tarefas diárias ainda representam um desafio e estão se tornando mais proibitivas, principalmente devido à grande quantidade de dados a serem processados. Portanto, técnicas de otimização eficientes são mais desejáveis que nunca.

Nesse contexto, a otimização meta-heurística ganhou destaque, sendo métodos estocásticos inspirados na natureza capazes de encontrar soluções sub-ótimas para problemas complexos com um esforço computacional razoável. No entanto, essas abordagens ainda sofrem com algumas desvantagens relacionadas à baixa convergência e estagnação em ótimos locais. Portanto, neste artigo, introduzimos conceitos da relatividade geral na conhecida técnica de otimização meta-heurística *Particle Swarm Optimization* (PSO). Os resultados experimentais validaram a robustez da abordagem proposta, que é comparada com o PSO padrão, além de outras três variações, em cinco funções de *benchmarking*.

Referência do trabalho:

RODER, M.; ROSA, G. H.; PASSOS JUNIOR, L. A.; PAPA, J. P.; ROSSI, A. L. D. Harnessing Particle Swarm Optimization Through Relativistic Velocity. *In: 2020 IEEE Congress on Evolutionary Computation*, Gasglow. 2020 IEEE Congress on Evolutionary Computation, p. 1-8, doi: 10.1109/CEC48606.2020.9185752, 2020.

8.6 *Energy-based Dropout in Restricted Boltzmann Machines: Why not go Random*

As arquiteturas de aprendizado profundo têm sido amplamente fomentadas nos últimos anos, sendo utilizadas em uma ampla gama de aplicações, como reconhecimento de objetos, reconstrução de imagens e processamento de sinais. No entanto, esses modelos sofrem de um

problema comum conhecido como *overfitting*, que limita a generalização da rede diante de dados desconhecidos de forma eficaz.

Abordagens de regularização surgem na tentativa de mitigar esse problema. Entre elas, pode-se citar o conhecido *Dropout*, que aborda o problema ao desligar aleatoriamente um conjunto de neurônios e suas conexões de acordo com uma determinada probabilidade. Entretanto, esta abordagem não considera nenhum conhecimento adicional para decidir quais unidades devem ser desconectadas. Neste artigo, propomos um *Dropout* baseado em energia (*E-Dropout*) que toma decisões conscientes acerca de quais neurônios devem ser desligados ou não.

Especificamente, projetamos este método de regularização correlacionando neurônios com a energia do modelo como um nível de importância, para aplicá-lo posteriormente em modelos baseados em energia, como Máquinas de Boltzmann Restritas (RBMs, do inglês *Restricted Boltzmann Machines*). Os resultados experimentais em vários conjuntos de dados de referência revelaram a adequação da abordagem proposta em comparação com o *Dropout* tradicional e as RBMs tradicionais.

Referência do trabalho:

RODER, M.; ROSA, G.; de ALBUQUERQUE, V. H. C.; ROSSI, A. L. D.; PAPA, J. P. Energy-based Dropout in Restricted Boltzmann Machines: Why not go random. *IEEE Transactions on Emerging Topics in Computational Intelligence*, p. 1-11, doi: 10.1109/TETCI.2020.3043764, 2020.

8.7 *On the Assessment of Nature-Inspired Meta-Heuristic Optimization Techniques to Fine-Tune Deep Belief Networks*

As técnicas de aprendizado de máquina são capazes de falar, interpretar, criar e até mesmo raciocinar sobre praticamente qualquer assunto. Além disso, seu poder de aprendizado cresceu exponencialmente nos últimos anos devido aos avanços nas arquiteturas de hardware. No entanto, a maioria desses modelos ainda têm dificuldades quanto ao seu uso prático, uma vez que requerem uma seleção adequada de hiper-parâmetros, que muitas vezes são escolhidos empiricamente.

Tais requisitos são reforçados quando se trata de modelos de aprendizado profundo, que comumente requerem um número maior de hiper-parâmetros. Uma coleção de técnicas de otimização inspiradas na natureza, conhecidas como meta-heurísticas, surgem como soluções diretas para lidar com esses problemas, uma vez que não empregam derivadas, aliviando assim sua carga computacional.

Portanto, este trabalho propõe uma comparação entre várias técnicas de otimização meta-heurísticas no contexto de ajuste fino de hiper-parâmetros de redes de crença profunda.

Uma configuração experimental foi conduzida em três conjuntos de dados públicos na tarefa de reconstrução de imagens binária e demonstrou resultados consistentes, apresentando técnicas meta-heurísticas como uma alternativa adequada para o problema.

Referência do trabalho:

PASSOS, L. A.; ROSA, G. H.; RODRIGUES, D.; RODER, M.; PAPA, J. P. On the Assessment of Nature-Inspired Meta-Heuristic Optimization Techniques to Fine-Tune Deep Belief Networks. *In: Iba H., Noman N. (eds) Deep Neural Evolution*. Natural Computing Series. Springer, Singapore. doi: 10.1007/978-981-15-3685-4_3, 2020.

8.8 *Enhancing Anomaly Detection Through Restricted Boltzmann Machine Features Projection*

A tecnologia vem alimentando uma ampla gama de aplicações nas últimas décadas, ajudando a automatizar algumas tarefas diárias. No entanto, sistemas de tecnologia mais avançada também expõem algumas falhas em potencial, que incentivam usuários mal-intencionados a explorar e tentar violar sua segurança.

Os pesquisadores tentaram superar esses problemas promovendo o chamado Sistema de Detecção de Intrusões, que são camadas de segurança que tentam detectar tentativas maliciosas (anomalias) e desativá-las. Além disso, o aumento da demanda por sistemas de aprendizado de máquina também permitiu a possibilidade de combinar essas abordagens, a fim de fornecer sistemas de detecção mais robustos.

Nesse contexto, foi introduzida uma nova abordagem para lidar com a detecção de anomalias, onde, em vez de usar os atributos brutos do problema, estes são projetados usando uma abordagem de codificação-decodificação (*encode-decode* por meio de *Restricted Boltzmann Machines*. A abordagem mencionada foi avaliada em um conjunto de dados de detecção de anomalias da literatura bem conhecido e obteve bons resultados, mostrando ser uma ferramenta promissora na projeção de atributos para detecção de anomalias.

Referência do trabalho:

ROSA, G. H.; RODER, M.; SANTOS, D. F. S.; COSTA, K. A. P. Enhancing anomaly detection through restricted Boltzmann machine features projection, *International Journal of Information Technology*, doi: 10.1007/s41870-020-00535-4, 2020.

8.9 *MaxDropout: Deep Neural Network Regularization Based on Maximum Output Values*

Diferentes técnicas surgiram no cenário de aprendizado profundo, como Redes Neurais Convolucionais, Redes de Crença em Profundidade e Redes de Memória de Longo Prazo, por exemplo. Em contrapartida, os métodos de regularização, que visam evitar o sobreajuste (do inglês, *overfitting*) penalizando algumas conexões, ou desligando neurônios, também têm sido amplamente estudados.

Neste artigo, apresentamos uma nova abordagem chamada *MaxDropout*, um regularizador para modelos de redes neurais profundas que funciona de forma supervisionada. Em suma, o método realiza o desligando de neurônios que são ativados a valores muito altos frente aos demais de uma dada camada oculta. Deste modo, os proeminentes são momentaneamente desligados essas conexões são forçadas a não possuírem valores absolutos tão elevados.

Com isso, o método força unidades menos ativas a aprender informações mais representativas, fornecendo assim esparsidade ao modelo final. Em relação aos experimentos, mostramos que é possível melhorar as redes neurais existentes e fornecer melhores resultados nestas quando o *Dropout* padrão é substituído pelo *MaxDropout*. O método proposto foi avaliado para a tarefa de classificação de imagens, obtendo resultados comparáveis aos regularizadores existentes, como *Cutout* e *Random-Erasing*.

Este trabalho foi aceito para publicação na conferência *25th Internation Conference on Pattern Recognition* - ICPR (2020), como co-autor.

Referências

ACKLEY, D.; HINTON, G.; SEJNOWSKI, T. J. A learning algorithm for boltzmann machines. In: WALTZ, D.; FELDMAN, J. (Ed.). *Connectionist Models and Their Implications: Readings from Cognitive Science*. Norwood, NJ, USA: Ablex Publishing Corp., 1988. p. 285–307. ISBN 0-89391-456-8.

AFONSO, B. et al. Moving-camera video surveillance in cluttered environments using deep features. In: . [S.l.: s.n.], 2018. p. 2296–2300.

BISHOP, C. *Neural networks for pattern recognition*. [S.l.]: Oxford University Press, 1995.

BOBICK, A. F. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical transactions of the Royal Society of London. Series B*, Biological sciences, v. 352, p. 1257–1265, 1997.

BRUZZONE, L.; MARCONCINI, M. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 32, n. 5, p. 770–787, 2009.

CARREIRA, J. et al. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

CONNOLLY, C. I. Learning to recognize complex actions using conditional random fields. In: SPRINGER. *International Symposium on Visual Computing*. [S.l.], 2007. p. 340–348.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

FEICHTENHOFER, C.; PINZ, A.; ZISSERMAN, A. Convolutional two-stream network fusion for video action recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 1933–1941.

FLEET, D.; WEISS, Y. Optical flow estimation. In: _____. *Handbook of Mathematical Models in Computer Vision*. [S.l.]: Springer US, 2006. p. 237–257.

GORELICK, L. et al. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 29, n. 12, p. 2247–2253, 2007.

GOWDA, S. N. Human activity recognition using combinatorial deep belief networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2017. p. 1–6.

HINTON, G. Training products of experts by minimizing contrastive divergence. *Neural Computation*, MIT Press, Cambridge, MA, USA, v. 14, n. 8, p. 1771–1800, 2002. ISSN 0899-7667.

HINTON, G. A practical guide to training restricted boltzmann machines. In: MONTAVON, G.; ORR, G.; MÜLLER, K.-R. (Ed.). *Neural Networks: Tricks of the Trade*. [S.l.]: Springer Berlin Heidelberg, 2012, (Lecture Notes in Computer Science, v. 7700). p. 599–619.

HINTON, G. E. Learning distributed representations of concepts. In: AMHERST, MA. *Proceedings of the eighth annual conference of the cognitive science society*. [S.l.], 1986. v. 1, p. 12.

HINTON, G. E. et al. The "wake-sleep"algorithm for unsupervised neural networks. *Science*, v. 268 5214, p. 1158–61, 1995.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. *Neural Computation*, MIT Press, Cambridge, MA, USA, v. 18, n. 7, p. 1527–1554, 2006. ISSN 0899-7667.

JIANG, Y.-G. et al. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, v. 2, n. 2, p. 73–101, 2013. ISSN 2192-662X.

JIANG, Y.-G. et al. *THUMOS challenge: Action recognition with a large number of classes*. 2014.

JIANG, Y.-G. et al. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. [S.l.: s.n.], 2011. p. 1–8.

KARPATHY, A. et al. Large-scale video classification with convolutional neural networks. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2014.

KHOJASTEH, P. et al. Exudate detection in fundus images using deeply-learnable features. *Computers in biology and medicine*, Elsevier, v. 104, p. 62–69, 2019.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR*. [S.l.: s.n.], 2015.

KUEHNE, H. et al. Hmdb: a large video database for human motion recognition. In: IEEE. *2011 International Conference on Computer Vision*. [S.l.], 2011. p. 2556–2563.

LAROCHELLE, H.; BENGIO, Y. Classification using discriminative restricted boltzmann machines. In: ACM. *Proceedings of the 25th international conference on Machine learning*. [S.l.], 2008. p. 536–543.

LECUN, Y. *Modèles connexionnistes de l'apprentissage*. Tese (Doutorado) — Paris 6, 1987.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998.

LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. Convolutional networks and applications in vision. In: *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. [S.l.: s.n.], 2010. p. 253–256.

LI, G.; HOU, Y.; WU, A. Fourth industrial revolution: technological drivers, impacts and coping methods. *Chinese Geographical Science*, v. 27, n. 4, p. 626–637, 2017. ISSN 1993-064X.

LI, W.; ZHANG, Z.; LIU, Z. Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE transactions on Circuits and Systems for Video Technology*, IEEE, v. 18, n. 11, p. 1499–1510, 2008.

LIU, J.; KUIPERS, B.; SAVARESE, S. Recognizing human actions by attributes. In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.]: IEEE Computer Society, 2011. (CVPR '11), p. 3337–3344. ISBN 978-1-4577-0394-2.

LIU, Y. et al. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, IEEE, v. 29, p. 3168–3182, 2019.

LOUI, A. et al. Kodak's consumer video benchmark data set: concept definition and annotation. In: *Proceedings of the international workshop on Workshop on multimedia information retrieval*. [S.l.: s.n.], 2007. p. 245–254.

MAHADEVAN, V. et al. Anomaly detection in crowded scenes. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2010. p. 1975–1981.

MARSZALEK, M.; LAPTEV, I.; SCHMID, C. Actions in context. In: IEEE. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.], 2009. p. 2929–2936.

MCCLELLAND, J. L.; RUMELHART, D. E.; HINTON, G. E. The appeal of parallel distributed processing. *MIT Press, Cambridge MA*, p. 3–44, 1986.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

MOHAMMADI, S. et al. Angry crowds: Detecting violent events in videos. In: LEIBE, B. et al. (Ed.). *Computer Vision – ECCV 2016*. [S.l.]: Springer International Publishing, 2016. p. 3–18. ISBN 978-3-319-46478-7.

NATARAJAN, P.; NEVATIA, R. Online, real-time tracking and recognition of human actions. In: IEEE. *2008 IEEE Workshop on Motion and video Computing*. [S.l.], 2008. p. 1–8.

NG, J. Y.-H. et al. Beyond short snippets: Deep networks for video classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 4694–4702.

NIEBLES, J. C.; CHEN, C.-W.; FEI-FEI, L. Modeling temporal structure of decomposable motion segments for activity classification. In: SPRINGER. *European conference on computer vision*. [S.l.], 2010. p. 392–405.

PASSOS, L. A.; PAPA, J. P. *On the training algorithms for Restricted Boltzmann Machine-Based Models*. Tese (Doutorado) — Universidade Federal de São Carlos, 2018.

PASSOS, L. A. et al. κ-entropy based restricted boltzmann machines. In: IEEE. *The 2019 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2019. p. 1–8.

QUATTONI, A.; COLLINS, M.; DARRELL, T. Transfer learning for image classification with sparse prototype representations. In: IEEE. *2008 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.], 2008. p. 1–8.

RAINA, R. et al. Self-taught learning: Transfer learning from unlabeled data. In: *Proceedings of the 24th International Conference on Machine Learning*. [S.l.]: Association for Computing Machinery, 2007. (ICML '07), p. 759–766. ISBN 9781595937933.

RAINA, R.; NG, A. Y.; KOLLER, D. Constructing informative priors using transfer learning. In: *Proceedings of the 23rd International Conference on Machine Learning*. [S.l.]: Association for Computing Machinery, 2006. (ICML '06), p. 713–720. ISBN 1595933832.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Cognitive modeling*, v. 5, n. 3, p. 1, 1986.

RUMELHART, D. E. et al. *Parallel distributed processing*. [S.l.]: IEEE, 1988. v. 1.

SALAKHUTDINOV, R.; HINTON, G. E. Deep boltzmann machines. In: *AISTATS*. [S.l.: s.n.], 2009. v. 1, p. 3.

SALAKHUTDINOV, R.; HINTON, G. E. An efficient learning procedure for deep boltzmann machines. *Neural Computation*, v. 24, n. 8, p. 1967–2006, 2012.

SALAKHUTDINOV, R.; MNIH, A.; HINTON, G. Restricted boltzmann machines for collaborative filtering. In: ACM. *Proceedings of the 24th international conference on Machine learning*. [S.l.], 2007. p. 791–798.

SCHULDT, C.; LAPTEV, I.; CAPUTO, B. Recognizing human actions: a local svm approach. In: IEEE. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* [S.l.], 2004. v. 3, p. 32–36.

SIMONYAN, K.; ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2014. p. 568–576.

SOOMRO, K.; ZAMIR, A. R.; SHAH, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

SUN, S.; SHI, H.; WU, Y. A survey of multi-source domain adaptation. *Information Fusion*, Elsevier, v. 24, p. 84–92, 2015.

TAS, Y.; KONIUSZ, P. CNN-based action recognition and supervised domain adaptation on 3D body skeletons via kernel feature maps. *arXiv preprint arXiv:1806.09078*, 2018.

ULLAH, A. et al. Action recognition using optimized deep autoencoder and cnn for surveillance data streams of non-stationary environments. *Future Generation Computer Systems*, Elsevier, v. 96, p. 386–397, 2019.

VAIL, D. L.; VELOSO, M. M.; LAFFERTY, J. D. Conditional random fields for activity recognition. In: ACM. *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. [S.l.], 2007. p. 235.

VENKATESWARA, H.; CHAKRABORTY, S.; PANCHANATHAN, S. Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations. *IEEE Signal Processing Magazine*, IEEE, v. 34, n. 6, p. 117–129, 2017.

WELLING, M.; ROSEN-ZVI, M.; HINTON, G. Exponential family harmoniums with an application to information retrieval. In: SAUL, L.; WEISS, Y.; BOTTOU, L. (Ed.). *Advances in Neural Information Processing Systems 17*. [S.l.]: MIT Press, 2005. p. 1481–1488.

ZHU, Y. et al. Hidden two-stream convolutional networks for action recognition. In: SPRINGER. *Asian Conference on Computer Vision*. [S.l.], 2018. p. 363–378.