UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

DETECÇÃO DE PORNOGRAFIA INFANTIL EM IMAGENS ATRAVÉS DE TÉCNICAS DE APRENDIZADO PROFUNDO

PAULO ROBERTO ROCHA VITORINO

ORIENTADOR: ANDERSON DE REZENDE ROCHA
CO-ORIENTADORA: SANDRA ELIZA FONTES DE AVILA

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
ÁREA DE CONCENTRAÇÃO INFORMÁTICA FORENSE E SEGURANÇA DA INFORMAÇÃO

PUBLICAÇÃO: PPGENE.DM - 621/16

BRASÍLIA / DF: DEZEMBRO/2016

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

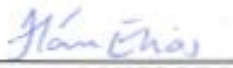
DETECÇÃO DE PORNOGRAFIA INFANTIL EM IMAGENS ATRAVÉS DE TÉCNICAS DE APRENDIZADO PROFUNDO

PAULO ROBERTO ROCHA VITORINO

DISSERTAÇÃO DE MESTRADO PROFISSIONAL SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE.

APROVADA POR:

ANDERSON DE REZENDE ROCHA, Dr., IC/UNICAMP
(ORIENTADOR)



FLÁVIO ELIAS GOMES DE DEUS, Dr., ENE/UNB
(EXAMINADOR INTERNO)

SIOVANI CINTRA FELIPUSSI, Dr., UFSCAR
(EXAMINADOR EXTERNO)

Brasília, 14 de dezembro de 2016.

FICHA CATALOGRÁFICA

VITORINO, PAULO ROBERTO ROCHA
Detecção de pornografia infantil em imagens através de técnicas de
aprendizado profundo [Distrito Federal] 2016.
xxii, 41p., 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2016).

Dissertação de Mestrado – Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica.

1. Reconhecimento de padrões 2. Visão por computador
3. Redes Neurais

I. ENE/FT/UnB. II. Título (Série)

REFERÊNCIA BIBLIOGRÁFICA

VITORINO, P. R. R. (2016). DETECÇÃO DE PORNOGRAFIA INFANTIL EM IMAGENS ATRAVÉS DE TÉCNICAS DE APRENDIZADO PROFUNDO. Dissertação de Mestrado, Publicação PPGENE.DM - 621/16, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 41p.

CESSÃO DE DIREITOS

NOME DO AUTOR: Paulo Roberto Rocha Vitorino
TÍTULO DA DISSERTAÇÃO: Detecção de pornografia infantil em imagens através de técnicas de aprendizado profundo.
GRAU/ANO: Mestre/2016.

Paulo Roberto Rocha Vitorino
Universidade de Brasília – Campus Universitário Darcy Ribeiro – Asa Norte
CEP 70910-900 – Brasília – DF - Brasil

Ao “Seu João” – *in memoriam* – grande amigo, e estimado pai. Saudades.

AGRADECIMENTOS

Aos meus pais, por me ensinarem os valores que realmente importam na formação do caráter, na compreensão do certo e do errado, e por sempre me incentivarem nos estudos.

Aos gestores da Delegacia de Polícia Federal em Guaíra/PR, DPF Reginaldo e DPF Smith, por entenderem a importância desta pesquisa e por darem o apoio necessário para a participação e conclusão do curso.

Ao responsável pela Unidade Técnico-Científica da Delegacia de Polícia Federal em Guaíra/PR, PCF Etienne, pelo apoio e incentivo, e pela compreensão, preocupação e amizade, que estavam sempre presente nas suas designações, e aos demais amigos desta Unidade, por suportarem a carga extra de trabalho que minhas ausências possam ter causado.

Aos amigos e companheiros de sala de aula, Cirilo Max, Egberto, Gustavo Henrique e Vitor, pela amizade, troca de conhecimento e por ajudarem a deixar mais suportável a semana de aulas longe da família.

Ao meu orientador, Anderson, e à minha co-orientadora, Sandra, pelo apoio, dedicação, paciência, compreensão e incentivo, sem os quais este trabalho não estaria concluído.

Aos meus filhos, Gabriel, Júlia, Amanda e Ana Beatriz, por serem a minha inspiração.

À minha esposa Adriana, que me deu todo apoio, incentivo e suporte durante esta jornada, e por sempre estar ao meu lado, nos bons, e nos maus momentos.

O presente trabalho foi realizado com o apoio da Polícia Federal – PF, com recursos do Programa Nacional de Segurança Pública com Cidadania – PRONASCI, do Ministério da Justiça e Cidadania.

RESUMO

DETECÇÃO DE PORNOGRAFIA INFANTIL EM IMAGENS ATRAVÉS DE TÉCNICAS DE APRENDIZADO PROFUNDO

Autor: Paulo Roberto Rocha Vitorino
Orientador: Anderson de Rezende Rocha
Programa de Pós-graduação em Engenharia Elétrica
Brasília, dezembro de 2016

Este trabalho apresenta uma nova abordagem para detecção de automática de pornográfica infantil em imagens, que se utiliza de técnicas de aprendizado profundo para extração das características discriminadoras de imagens, e um classificador de padrões baseado em máquinas de vetores de suporte, para determinar se as imagens contêm, ou não, pornografia infantil (PI). Adicionalmente, também é proposta técnica baseada em sacolas de palavras para resolver o problema. As soluções desenvolvidas atingem um acerto de +87% de acurácia de classificação quando separando conteúdo de pornografia infantil de conteúdos de pornografia geral e imagens normais, sobressaindo-se em relação às técnicas existentes na literatura.

ABSTRACT

CHILD PORNOGRAPHY IMAGE DETECTION THROUGH DEEP LEARNING TECHNIQUES

Author: Paulo Roberto Rocha Vitorino
Supervisor: Anderson de Rezende Rocha
Programa de Pós-graduação em Engenharia Elétrica
Brasília, December of 2016

In this work, we present a new method for automatic detection of sexually exploitative imagery of children (SEIC) or child pornography content. Our solution leverages cutting-edge concepts of deep learning – for extracting discriminative features from images – and the support vector machine classifier, it point out whether or not an image contains child pornography content. Moreover, it is also proposed one technique based on bags of visual words methodology to deal with this difficult problem. The developed solutions lead to as much as 87% classification accuracy when separating SEIC content from adult (adult pornography) and other seemingly innocuous content (everyday image content) clearly outperforming existing counterparts in the literature.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SÍMBOLOS, NOMENCLATURA E ABREVIAÇÕES

BoVW *Bag of Visual Words*

CNN *Convolution Neural Network*

CP *Child Pornography*

CSA *Child Sexual Abuse*

CVIP *Child Victim Identification Program*

DNN *Deep Neural Network*

NCMEC *National Center for Missing and Exploited Children*

PCA *Principal Component Analysis*

PI Pornografia Infantil

SVM *Support Vector Machine*

1. INTRODUÇÃO

Neste capítulo, são apresentados dados referentes ao problema que motivou este trabalho, alguns trabalhos relacionados, bem como o objetivo e as contribuições do mesmo.

1.1. MOTIVAÇÃO

O avanço tecnológico visto nas últimas décadas nos trouxe a uma convergência na transformação da informação, dos mais variados meios, ao meio digital. Aliado a essa convergência, temos mais do que nunca, facilidade em compartilhar conteúdos a partir de um clique. Dispositivos móveis, como os *smartphones*, já são capazes de produzir em meio digital, informação como imagem ou vídeo, e, em uma fração de segundo, disponibilizá-la em redes sociais, via Internet.

Consequentemente, a utilização de todo esse aparato tecnológico para prática de crimes tem sido cada vez mais comum, o que vem demandando dos órgãos de criminalística a capacidade de processar e analisar maior quantidade e volume de evidências digitais do que há poucos anos. Dentre os crimes praticados por meio de computadores, a produção, distribuição e consumo de material multimídia relacionado à exploração sexual de crianças e adolescentes tem se destacado, não só pela facilidade com que se produz e se dissemina esse conteúdo em meio digital, mas, principalmente, pelo constante aumento na disponibilização desse tipo de material.

Segundo o *National Center for Missing and Exploited Children* – NCMEC[1], no período de 2005 a 2011, houve um incremento de 774% no número de imagens e vídeos, contendo pornografia infantil, verificadas pelos analistas do programa de identificação de vítimas (*Child Victim Identification Program* – CVIP). No ano de 2013, o CVIP analisou 22 milhões de arquivos de imagem/vídeo contendo pornografia infantil – aumento de mais de 5.000% em relação a 2007. Até abril de 2015, os analistas do CVIP já haviam verificado mais de 139 milhões de imagens e vídeos com tal conteúdo.

No período de 2012 a 2015, de acordo com dados obtidos do Sistema Nacional de Gestão da Criminalística da Polícia Federal do Brasil, foram realizados exames periciais em mais 4.000 dispositivos de armazenamento computacional, superando 500 TB de dados analisados.

[1] http://www.missingkids.com/

Infelizmente, verifica-se, no dia-a-dia, que o volume de material produzido e a quantidade de meios através dos quais este material é distribuído é muito maior que a capacidade de análise visual feita por profissionais das forças da lei.

1.2. TRABALHOS CORRELATOS

Pelo exposto, a classificação de conteúdo nessas mídias, automática e de forma inteligente e contínua, é de suma importância, não só para aplicação em métodos relacionados a pesquisas de imagens e vídeos, mas, principalmente, para reconhecimento de conteúdos que podem ser considerados indesejados ou ofensivos, a fim de ser capaz de detectar estes materiais.

Neste contexto, diversas metodologias já foram propostas para detecção automática de pornografia infantil em imagens e vídeos, seguindo abordagens já apresentadas para detecção automática de pornografia. Tipicamente, as soluções encontradas na literatura para a detecção automática de pornografia são baseadas em:

(1) detecção de pele humana (Fleck et al. 1996; Forsyth & Fleck 1999; Zheng et al. 2004);
(2) características da distribuição espacial para reconhecimento do corpo humano (Jones & Rehg 2002; Lee et al. 2009; Bouirouga et al. 2012);
(3) extração de características locais e modelos Sacolas de Palavras Visuais (Deselaers et al. 2008; Jansohn et al. 2009; Steel 2012; Avila et al. 2013); e
(4) técnicas de aprendizado profundo (Moustafa 2015; Perez 2016).

Adicionalmente, na detecção de pornografia infantil, uma abordagem muito difundida é a comparação de assinaturas únicas (*hash*) dos arquivos (Microsoft 2009; Oliveira & Silva 2009; Vrubel 2011).

Abordagens baseadas em (1) e (2) naturalmente sofrem com o alto número de falso positivos (quando imagens normais são consideradas como sendo de pornografia infantil) uma vez que há um grande fosso semântico entre o conceito de exposição de pele e a extrapolação de que isso está diretamente ligado à pornografia, pois nem todas as imagens com grandes áreas de exposição de pele são necessariamente pornográficas, como imagens com pessoas usando roupas de banho, ou imagens relacionadas a esportes.

Complementarmente, abordagens baseadas em (3) são mais robustas que as baseadas em detecção de pele, obtendo melhores resultados na classificação, mas ainda encontram dificuldade em casos ambíguos. De uma maneira geral, essas abordagens adotam um conceito multi-nível. No primeiro nível, as principais informações presentes em uma imagem de entrada são codificadas utilizando o conceito de descritores locais de baixo nível. Em seguida, tais descritores de baixo nível são mapeados em uma representação de médio nível utilizando o conceito de dicionários visuais. Finalmente, os conceitos representados no médio nível são capturados em uma representação semântica mais sofisticada (alto nível) para o conceito pornografia vs. não pornografia. Esse último passo é geralmente implementado a partir de um classificador de padrões como, por exemplo, baseado em máquinas de vetores de suporte (SVM).

Para a detecção de pornografia infantil, além da utilização das abordagens anteriores (Ulges & Stahl 2011; Carvalho 2012), foram propostas novas abordagens que combinaram algumas das anteriores (Polastro & Eleuterio 2010), ou que adicionaram a elas funcionalidades específicas, como o reconhecimento facial, para identificar imagens de crianças (Sae-Bae et al. 2014).

Mais recentemente, abordagens baseadas em (4) têm surgido, impulsionadas pelos excelentes resultados apresentados pela técnica de aprendizado profundo (*deep learning*), em especial as redes neurais convolucionais (*Convolutional Neural Networks* – CNN), nas tarefas de classificação de imagens e vídeos (Krizhevsky et al. 2012; Szegedy et al. 2015). As CNNs superaram as abordagens anteriores, com grande folga, em diferentes *datasets* e em desafios de classificação de imagens, como o ImageNet[2]. Esta é a abordagem escolhida para utilização neste trabalho.

1.3. OBJETIVO

Este trabalho tem como principal objeto propor uma nova abordagem para detecção automática de pornografia infantil em imagens, que se utiliza de técnicas de aprendizado profundo para extração de características discriminadoras de imagens, e um classificador de padrões baseado em máquinas de vetores de suporte, para determinar se as imagens contêm, ou não, pornografia infantil (PI). Adicionalmente, também é proposta uma técnica baseada em sacolas de palavras como um caminho para resolver o problema (Vitorino et al. 2016).

[2] http://www.image-net.org/

1.4. CONTRIBUIÇÕES

Dentre as principais contribuições deste trabalho, destaca-se a inovação na aplicação de técnicas de aprendizado profundo (*deep learning*) na tarefa de detecção automática de imagens de pornografia infantil, com a utilização do *framework* de código aberto Caffe[3] (Jia et al. 2014). Adicionalmente, a arquitetura treinada é reutilizável, criada a partir do processamento de dezenas de milhares de imagens reais de PI, originárias do Banco de Imagens de Pornografia Infantil do Setor Técnico-Científico da Superintendência da Polícia Federal no Paraná (SETEC/SR/PF/PR), instituído pela Instrução de Serviço (IS) 08/2012-SETEC/SR/DPF/PR, de 26/09/2012.

Devido à vedação legal (Brasil 2016), as imagens de PI foram manipuladas exclusivamente por agente público no exercício de suas funções, e por obrigatoriedade de sigilo, não poderão ilustrar este trabalho. No entanto, a rede neural aprendida para extrair características e classificar imagens quanto à presença de pornografia infantil pode ser utilizada por qualquer perito interessado em colocar nossa pesquisa em operação no campo.

1.5. ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado como descrito a seguir. No Capítulo 2, é apresentada a revisão do estado da arte, onde são discutidas as principais abordagens utilizadas para detecção automática de pornografia e de pornografia infantil. No Capítulo 3, são descritos os conceitos relacionados à abordagem adotada, necessários para melhor entendimento da metodologia proposta. O Capítulo 4 descreve a metodologia desenvolvida neste trabalho. O Capítulo 5 apresenta os experimentos e resultados obtidos com a aplicação proposta na detecção automática de PI. Por fim, o Capítulo 6 é dedicado às conclusões e trabalhos futuros.

[3] http://caffe.berkeleyvision.org/

2. ESTUDO BIBLIOGRÁFICO – ESTADO DA ARTE

Em um estudo recente, (Short et al. 2012) revisou metodologia e conteúdo de 44 artigos relativos a pornografia na Internet, e salientou a importância de que autores, em suas pesquisas, façam uma definição explícita do que é considerado pornografia, uma vez que isto tem relação direta com os resultados obtidos. Esta definição é importante, também, para comparação de trabalhos correlatos. Para exemplificar, alguns artigos consideram a exibição de genitálias suficiente para classificação de pornografia, enquanto que em outros se entende que é preciso o ato sexual explícito. Assim, também deve ser com a pornografia infantil (PI). A definição de PI adotada nesta pesquisa é a constante na legislação que criminaliza esse tipo de conteúdo, o Estatuto da Criança e do Adolescente (Brasil 2016):

> Art. 241-E. Para efeito dos crimes previstos nesta lei, a expressão "cena de sexo explícito ou pornográfica" compreende qualquer situação que envolva criança ou adolescente em atividades sexuais explícitas, reais ou simuladas, ou exibição dos órgãos genitais de uma criança ou adolescente para fins primordialmente sexuais.

Segundo (Digiácomo & Digiácomo 2013), com essa definição, o legislador quis evitar possíveis dúvidas quanto ao alcance da norma proibitiva, que deve ser o mais abrangente possível.

2.1. DETECÇÃO DE TONS DE PELE

Técnicas de detecção de tons de pele têm sido amplamente exploradas para detecção de nudez (Fleck et al. 1996; Forsyth & Fleck 1999). Uma estratégia para identificar imagens de pessoas nuas utilizando recuperação baseada em conteúdo foi proposta por (Fleck et al. 1996). Nela, são usados *thresholds* para os valores de intensidade, matiz e saturação de cada pixel, de forma a identificar um pixel de pele. Imagens com grandes áreas (regiões) de pele são então filtradas. Essas regiões são analisadas geometricamente buscando identificar se elas podem representar partes do corpo humano. Em (Jones & Rehg 2002) foram elaborados modelos estatísticos de tons de pele, focando exclusivamente nas informações de cor dos pixels. São gerados histogramas de 256 *bins* em cada canal, um para imagens de pele, e outro para imagens sem pele. Esses histogramas modelam a probabilidade do pixel ser de uma região de pele.

(Kovac et al. 2003) propuseram um detector de pele humana que utiliza testes lógicos para determinar a classe que a imagem pertence, a partir das diferenças entre os valores do pixel no espaço de cor RGB (c.f., Equação 2.1). Dessa forma, determina-se um subespaço de cores capaz de identificar a pele humana.

$$\begin{aligned}&\{[(\mathrm{R}>95)\wedge(\mathrm{G}>40)\wedge(\mathrm{B}>20)]\wedge\\&[\max(\mathrm{R},\mathrm{G},\mathrm{B})-\min(\mathrm{R},\mathrm{G},\mathrm{B})>15]\wedge\\&[|\mathrm{R}-\mathrm{G}|>15\wedge(\mathrm{R}>\mathrm{G})\wedge(\mathrm{R}>\mathrm{B})]\}.\end{aligned} \qquad \text{(Eq. 2.1)}$$

Na detecção de imagens de pornografia infantil (PI), técnicas de detecção de nudez são amplamente utilizadas. Nas abordagens propostas por (Polastro & Eleuterio 2010), (Islam et al. 2011), (Kawale & Patil 2014), (Schulze et al. 2014) e (Sae-Bae et al. 2014), diferentes métodos para detecção de pele são utilizados para filtrar imagens que possam conter nudez e, com a aplicação de outras técnicas, classificar essas imagens como sendo de PI.

2.2. SACOLA DE PALAVRAS VISUAIS

Diante dos bons resultados em diversos problemas de classificação de imagens, (Deselaers et al. 2008) propuseram a utilização da abordagem de sacola de palavras visuais (*bag-of-visual-words* – BoVW) para detecção de pornografia. Foram usados como características fragmentos centrados em pontos de interesse, após a redução de dimensionalidade via *Principal Component Analysis* (PCA). O *codebook* foi gerado a partir da seleção de palavras visuais (*codewords*) através de modelos de gaussianas. Para a classificação foram utilizados os classificadores Support Vector Machine (SVM) e Log-linear. Os resultados apresentados mostraram o desempenho superior deste método sobre outros, focados em características de cor, e que a combinação deste método com soluções baseadas em detecção de pele não representou um ganho considerável de desempenho.

A abordagem BoVW também já foi avaliada na detecção de pornografia infantil. (Ulges & Stahl 2011) apresentou uma abordagem baseada na extração de palavras visuais de características de cores (*color visual word features*), e classificação estatística com SVM. As palavras visuais foram extraídas com a aplicação de DCT (*Discrete Cosine Transform*) no espaço de cores YUV, e a seleção de 78 coeficientes de baixa-frequência (36 para iluminação

e 21 para canal cromático). A taxa de erro desta abordagem ficou entre 11 e 24%, conforme a classe negativa utilizada.

Com o objetivo de analisar o desempenho de extratores de palavras visuais na classificação de imagens de PI, (Carvalho 2012) realizou experimentos combinando diferentes detectores de pontos de interesse, descritores locais e classificadores. Foram utilizados, alternadamente, os detectores Harris-Laplace e Amostragem Densa com os descritores OpponentSIFT e WSIFT, e classificador Naive Bayes; e o detector Difference of Gaussians com o descritor SIFT e classificador pLSA. Segundo o autor, o detector Harris-Laplace se destacou nas classificações mais complexas, enquanto o descritor SIFT apresentou baixo desempenho nas comparações mais difíceis, por não utilizar informações de cores.

2.3. REDES NEURAIS CONVOLUCIONAIS

Atualmente, a maior parte dos avanços na classificação de imagens é creditada à abordagem de aprendizado profundo (*deep learning*), no entanto, poucos experimentos foram realizados utilizando essas técnicas no contexto da detecção de pornografia.

Adaptando superficialmente arquiteturas CNN (*Convolutional Neural Networks*) de destaque na classificação de imagens, (Moustafa 2015) usou as arquiteturas AlexNet (Krizhevsky et al. 2012) e GoogLeNet (Szegedy et al. 2015) em *key-frames* de vídeos, como imagens individuais, para classificação de pornografia em vídeos, após a votação majoritária desses frames.

Também adaptando a arquitetura proposta por (Krizhevsky et al. 2012), (Huang & Kong 2016) utilizaram um conjunto de CNNs para detecção de imagens de pornografia e de *upskirt*. Sete CNNs foram treinadas individualmente, em *datasets* balanceados com diferentes imagens de pornografia e imagens comuns, mas sempre com as mesmas imagens de *upskirt*. Essas CNNs foram então combinadas em conjuntos de *n* elementos, onde *n* varia de uma sete. Por fim, a acurácia média de todos os conjuntos de *n* elementos é reportada (ECNN-*n*). O conjunto de 6 CNNs (ECNN-6) obteve os melhores resultados, com acurácia de 91,28% na classificação de duas classes (pornografia e *upskit*), e de 90,23% na classificação de três classes (pornografia, *upskit* e normal).

Abordagem que vem sendo muito utilizada na detecção de pornografia infantil, a estimativa de idade também tem nas CNNs um futuro promissor. (Wang et al. 2015) construiu um modelo para extração de características de idade baseado em aprendizado profundo, avaliando diferentes abordagens de coletores de aprendizagem (*manifold learning*), e métodos de classificação e regressão, como *Support Vector Regression* (SVR), *Support Vector Machines* (SVM) e *Partial Least Squares* (PLS), para estimativa de idade. Utilizando dois *datasets* públicos para comparação com o estado da arte, o modelo proposto obteve o menor *Mean Absolute Error* (MAE) dentre as abordagens avaliadas, atingindo índices de 4,77 e 4,26 para os *datasets* MORPH e FG-NET, respectivamente.

Nossa revisão bibliográfica sugere que não foram divulgados, até o momento, estudos sobre a aplicação da abordagem de aprendizado profundo no problema de detecção de pornografia infantil.

3. CONCEITOS RELACIONADOS

Neste capítulo, são apresentados alguns conceitos necessários para a compreensão deste trabalho. A metodologia proposta está centrada na utilização de técnicas de aprendizado profundo (*deep learning*). Por esta razão, seus conceitos, como noções básicas de Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNN), aperfeiçoamento de filtros e arquitetura GoogLeNet (Szegedy et al. 2015) são apresentados na Seção 3.1. As técnicas associadas aos métodos de descrição de características locais, que usamos para comparação com a metodologia proposta, são apresentadas na Seção 3.2.

3.1. APRENDIZAGEM PROFUNDA

A principal razão para o sucesso das redes neurais profundas (DNNs, do inglês *Deep Neural Networks*) está na sua capacidade de extrair características de dados brutos, em especial, os valores de intensidade dos pixels. Essas características ajudam a diminuir o *gap* semântico, que consiste na distância entre o valor semântico de uma imagem, como, por exemplo, um objeto, e a representação desta imagem no computador, obtida pela análise de conteúdo de baixo nível (pixels). Dessa forma, sistemas de aprendizagem supervisionados, como o SVM, podem treinar com uma representação de mais alto nível, gerando modelos altamente eficazes.

Nesta seção, apresentamos as noções básicas da arquitetura abordada em nossa metodologia proposta, métodos de aperfeiçoamento e o projeto do GoogLeNet (Szegedy et al. 2015).

3.1.1. Redes Neurais Convolucionais

As redes neurais convolucionais inspiradas no processo biológico de processamentos de dados visuais. Neste processamento, é a associação de informações próximas, ou correlação espacial, é de suma importância.

As redes convolucionais reforçam este tipo de correlação ao conectar apenas um subconjunto de entradas aos neurônios de camadas adjacentes. Isso significa que cada neurônio estará conectado a apenas uma amostragem dos neurônios da camada inferior, e a saída desses neurônios está relacionada a regiões próximas uma da outra na entrada original como, por

exemplo, pixels próximos em uma imagem. A Figura 3.1 mostra uma representação visual desta conectividade esparsa entre as camadas de neurônios.

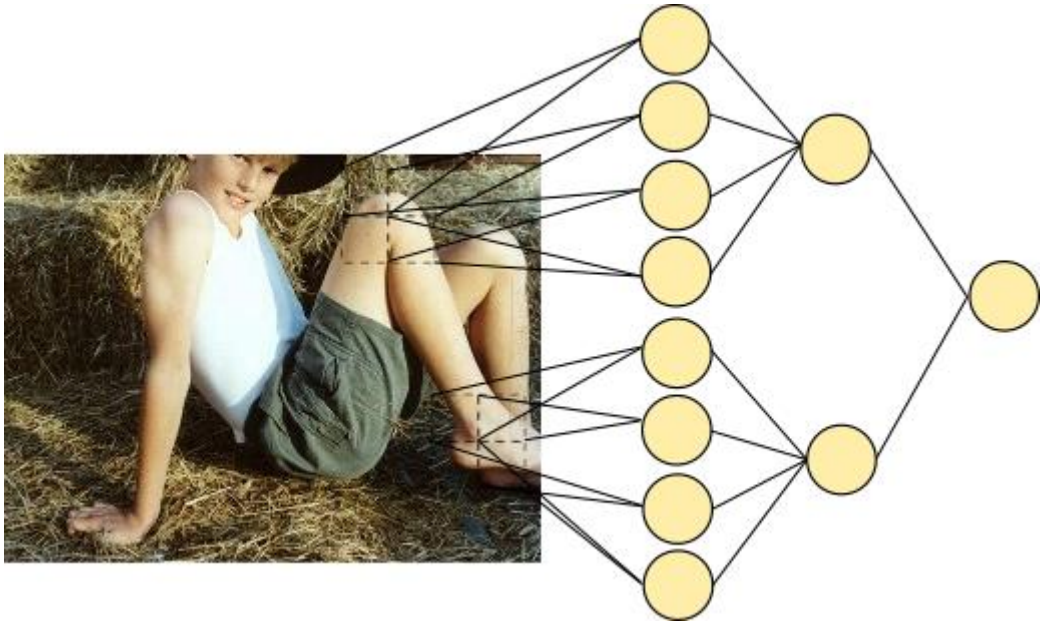


Figura 3.1 – Conectividade esparsa para reforçar a correlação espacial. Imagem baseada em (Perez 2016).

Dessa forma, ao agregarmos informações de áreas menores nas camadas inferiores, permitimos que a rede gere uma informação mais precisa, na camada superior, de uma área maior que a processada na camada anterior. As áreas que um neurônio recebe como entrada são chamadas de campos receptivos locais (*local receptive fields*). Assim, a pilha de camadas convolucionais vai da informação local, como dados de pixels do tipo cor e brilho, para conceitos mais amplos, tal como bordas e cantos. Através do aprendizado, esses conceitos evoluem naturalmente dos cantos para as noções de alto nível, como olhos e narizes (Zeiler & Fergus 2014).

O compartilhamento de parâmetros entre um subconjunto dos neurônios na mesma camada está presente nas redes convolucionais para que o mesmo tipo de informação possa ser filtrado quando presente em diferentes locais de uma imagem, já que os conceitos mencionados anteriormente podem aparecer várias vezes, e em regiões distintas. A coleção de parâmetros compartilhados por um subconjunto de neurônios de uma mesma camada define um mapa de características.

Como cada um dos neurônios deste subconjunto tem uma combinação distinta de entradas no campo receptivo, devido à conectividade esparsa, mas possui parâmetros em comum com outros neurônicos da camada, vemos que eles analisam áreas diferentes, mas em busca do mesmo conteúdo. A Figura 3.2 representa o uso comum de pesos entre os neurónios com um mesmo mapa de características.

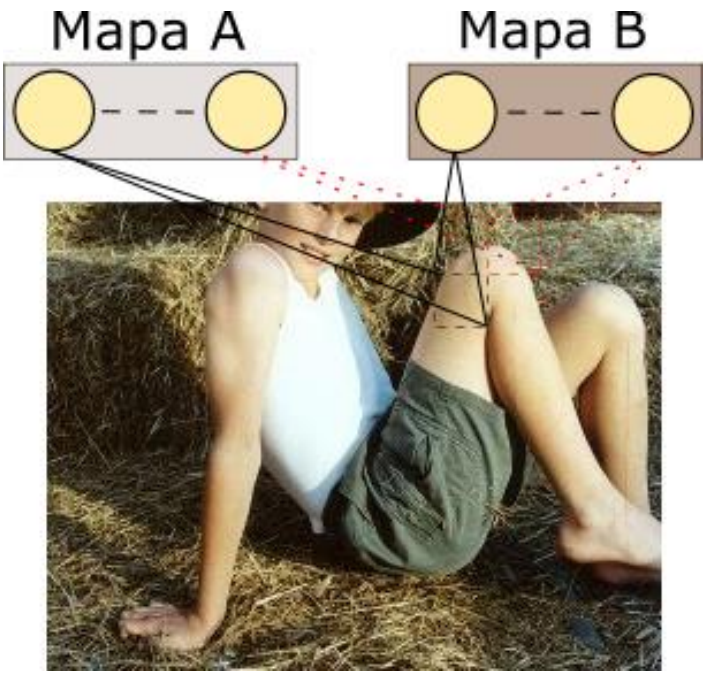


Figura 3.2 – Uso comum de pesos entre os neurónios com um mesmo mapa de características. Imagem baseada em (Perez 2016).

Os mapas de características mencionados anteriormente são, de fato, um conjunto de pesos, compondo um banco de filtros. Os bancos de filtros são aplicados ao sinal de entrada através de uma operação matemática denominada convolução, daí o nome desta arquitetura de rede. Para a aprendizagem dos pesos do filtro, normalmente aplicamos uma descida de gradiente (Russakovsky et al. 2015). O aperfeiçoamento dos filtros é um passo muito delicado e determinante para o desempenho da rede, razão pela qual ele será abordado detalhadamente na próxima subseção.

Usualmente, após aplicação do banco de filtros na entrada através da convolução, a saída é submetida a uma função de ativação não-linear. Após a aplicação do mapa de característica, áreas podem ter suas saídas agrupadas por *max-pooling*, antes de serem encaminhadas para a camada superior. Essa etapa é opcional, mas ajuda a reduzir a dimensionalidade e permite a composição da informação de invariância espacial.

Do exposto, podemos dizer que as redes neurais convolucionais consistem basicamente em empilhar camadas de convolução e agrupamento (*pooling*), geralmente mais de vinte, algumas camadas totalmente conectadas, e uma camada softmax para a classificação. Resumidamente, os parâmetros estruturais da rede convolucional básica podem ser resumidos em:

a. Quantidade de camadas;
b. Número de filtros em cada camada;
c. Tamanho e passo (*stride*) dos campos receptivos;
d. Tipo da função não-linear;
e. Tamanho e passo do agrupamento.

3.1.2. Aperfeiçoamento do filtro

Mesmo tendo uma rede neural convolucional de arquitetura adequada, se forem usados pesos aleatórios para seus filtros, o desempenho ficará aquém do desejado, impossibilitando a extração das melhores características das imagens, por exemplo. Para determinarmos os filtros de aprendizagem mais adequados a um determinado problema, empregamos uma técnica de descida de gradiente estocástico (*Stochastic Gradient Descent*), usando *backpropagation* para computar os gradientes (Russakovsky et al. 2015).

Em resumo, o aprendizado consiste em aperfeiçoar uma função, que está relacionada aos pesos da rede, analisando as derivadas (ou gradiente) dessa função em cada camada, “retro propagando” os gradientes da camada superior (saída de predição) para a camada inferior (entrada de dados). A partir de suas respectivas informações de gradiente, os pesos da rede sofrem pequenos ajustes, visando aperfeiçoar a função.

O aperfeiçoamento de filtros, especialmente em redes profundas (a partir de vinte camadas), pode ter um efeito secundário indesejado, também comum em outras abordagens de aprendizado de máquina, conhecido como *overfitting*. O *overfitting* está associado à forte especialização dos parâmetros, aperfeiçoados durante o treinamento, sobre os dados de treinamento, deixando-os sem a generalidade necessária para lidar com os dados de teste, ainda não vistos. Isso acaba resultando em um alto desempenho no treinamento, mas um mau desempenho durante o teste. Esta anomalia acontece com frequência quando se utiliza DNNs, principalmente porque elas contêm um grande número de parâmetros para modelagem das amostras alimentadas durante o processo de aprendizagem.

O *overfitting* pode ser mitigado, ou até mesmo evitado, se for utilizada alguma das contramedidas existentes. Uma abordagem comum para combater o *overfitting* é utilizar mais dados durante o treinamento, o que nem sempre é possível. Um método alternativo que tem sido muito aplicado (Krizhevsky et al. 2012; Russakovsky et al. 2015), chamado de aumentação dos dados (*data augmentation*), consiste em criar amostras adicionais, a partir dos dados existentes. O método consiste em aplicar manipulações básicas nas imagens, como redimensionar, cortar regiões diferentes ou rotacionar, para aumentar o número de amostras de treinamento. Embora técnicas como aumentação tenham sido desenvolvidas para um cenário de dados de treinamento insuficientes, elas podem ser aplicadas em qualquer caso, visando melhorar o desempenho de um determinado sistema de classificação.

Do exposto nesta subseção, podemos observar que o aperfeiçoamento de filtros, além de difícil, é de suma importância para se alcançar um desempenho satisfatório durante a concepção e o desenvolvimento de uma rede neural convolucional profunda. Felizmente, filtros aperfeiçoados de um treinamento anterior podem ser usados como ponto de partida para aprender com novos dados de treinamento, de um novo problema, desde que haja alguma relação entre eles. Esta técnica é referenciada na literatura como transferência de conhecimento e, posteriormente, ajuste de pesos (*fine-tuning*).

Assim, através do *fine-tuning*, podemos transferir o aprendizado de outro conjunto de dados, com muitas amostras, para o nosso problema. Dessa forma, diminuimos a necessidade de um grande número de amostras de dados de treinamento para o problema-alvo. Isso normalmente leva a um desempenho melhor do que aquele obtido quando se treina a partir do zero, usando apenas os dados existentes (normalmente em baixa quantidade) do problema-alvo.

Na metodologia proposta neste trabalho, apresentada na Seção 4, são empregados os conceitos de aperfeiçoamento de filtros explanados anteriormente. Aplica-se o *fine-tuning* treinando a arquitetura GoogLeNet (Szegedy et al. 2015) a partir dos pesos aprendidos com imagens de pornografia comum. Neste caso, o modelo GoogLeNet (Szegedy et al. 2015) também foi treinado utilizando a técnica *data augmentation*, objetivando melhorar o desempenho do sistema de classificação. O *fine-tuning*, na verdade, começa com a arquitetura GoogLeNet (Szegedy et al. 2015) e os pesos aprendidos do Imagenet 2014 em mais de um milhão de imagens naturais (Russakovsky et al. 2015). A partir desses pesos e desse problema-fonte (classificação de imagens naturais), fazemos a primeira transferência de conhecimento para um problema-alvo (detecção de pornografia). Posteriormente, transferimos novamente o conhecimento do problema de detecção de pornografia (agora considerado

problema-fonte) para o novo problema-alvo (detecção de pornografia infantil). Esses dois níveis de transferência de conhecimento e adaptação de pesos foi essencial para conseguirmos desenvolver uma solução eficaz para o problema proposto.

3.1.3. GoogLeNet

Diferentes modelos de redes neurais profundas têm sido propostos e avaliados ao longo dos últimos anos. Mas foi somente na edição de 2012 do desafio ImageNet (Russakovsky et al. 2015), com a contribuição de (Krizhevsky et al. 2012), que a comunidade de pesquisa realmente voltou sua atenção para a aprendizagem profunda.

A rede neural convolucional que os autores construíram para a competição ficou em primeiro lugar, melhorando a marca do segundo colocado em mais de dez pontos percentuais, representando uma melhora significativa nos problemas de classificação de imagens. Desde então, houve grandes melhorias na arquitetura das CNNs e técnicas de treinamento. Na edição de 2014 do ImageNet, (Szegedy et al. 2015) apresentaram a GoogLeNet, uma CNN que obteve uma melhora da precisão de classificação bastante significativo em relação aos resultados de (Krizhevsky et al. 2012).

3.1.3.1. Arquitetura

Os principais componentes da arquitetura GoogLeNet (Szegedy et al. 2015) são os módulos iniciais. Este módulo executa 1x1, 3x3 e 5x5 convoluções, mais um *max-pooling* de 3x3 na mesma camada convolucional, e em cima da mesma entrada, concatenando todas as saídas em um único vetor. Devido a requisitos computacionais, os bancos de filtro de 1x1 são aplicados antes das convoluções de 3x3 e 5x5, e após o *max-pooling* de 3x3, para a redução dimensional da entrada antes destas operações mais caras. A Figura 3.3 mostra um módulo inicial.

Assim, a GoogLeNet (Szegedy et al. 2015), que recebe como entrada uma imagem RGB de dimensões 224x224 pixels, é construída como uma pilha de módulos iniciais, juntamente com outras camadas de uso comum em redes neurais convolucionais, tais como convoluções regulares, totalmente conectadas, *max-pooling, dropout* e softmax. A função não-linear é usada em todas as camadas de convolução, redução e *pooling*. Se considerarmos as camadas de *pooling*, a arquitetura é formada por um total de vinte e sete camadas.

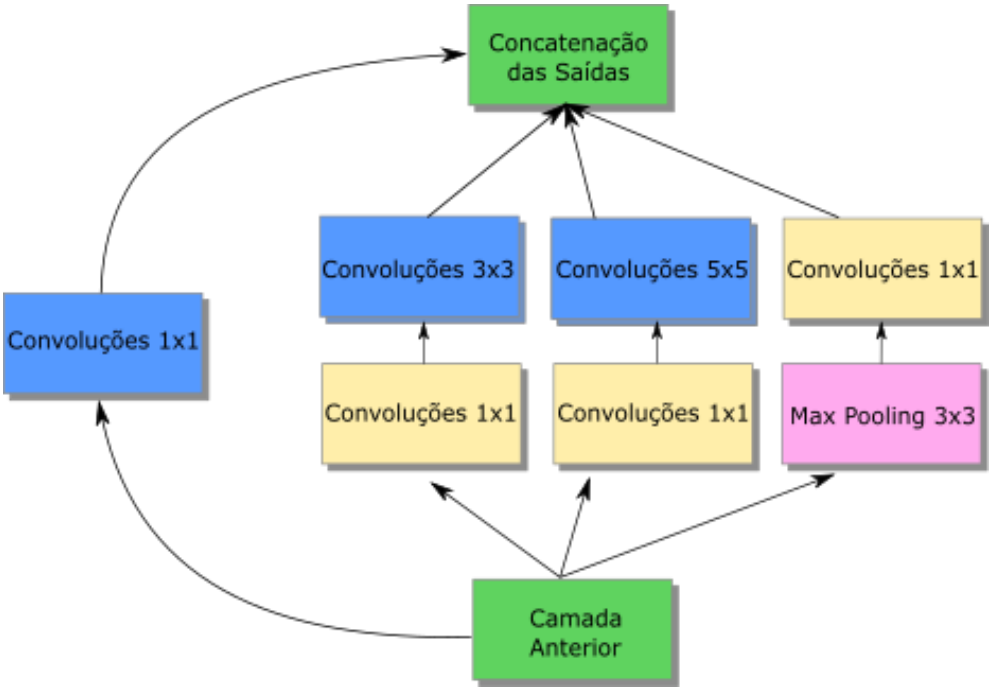


Figura 3.3 – Representação de um módulo inicial presente na arquitetura GoogLeNet. Imagem baseada em (Szegedy et al. 2015).

Além do classificador final, a rede também contém dois classificadores auxiliares. Estes classificadores foram inseridos em camadas intermediárias como forma de mitigar o problema do gradiente de fuga (Hochreiter et al. 2000), que ocorre devido à elevada profundidade da rede. Os classificadores extras são acrescentados ao topo de módulos iniciais selecionados e tomam forma como CNNs menores, e são estruturados com os seguintes componentes:

1) Camada de *average pooling*;
2) Camada de redução de convolução 1x1;
3) Camada totalmente conectada;
4) Camada *dropout*;
5) Softmax como classificador.

A Figura 3.4 mostra toda a arquitetura do GoogLeNet.

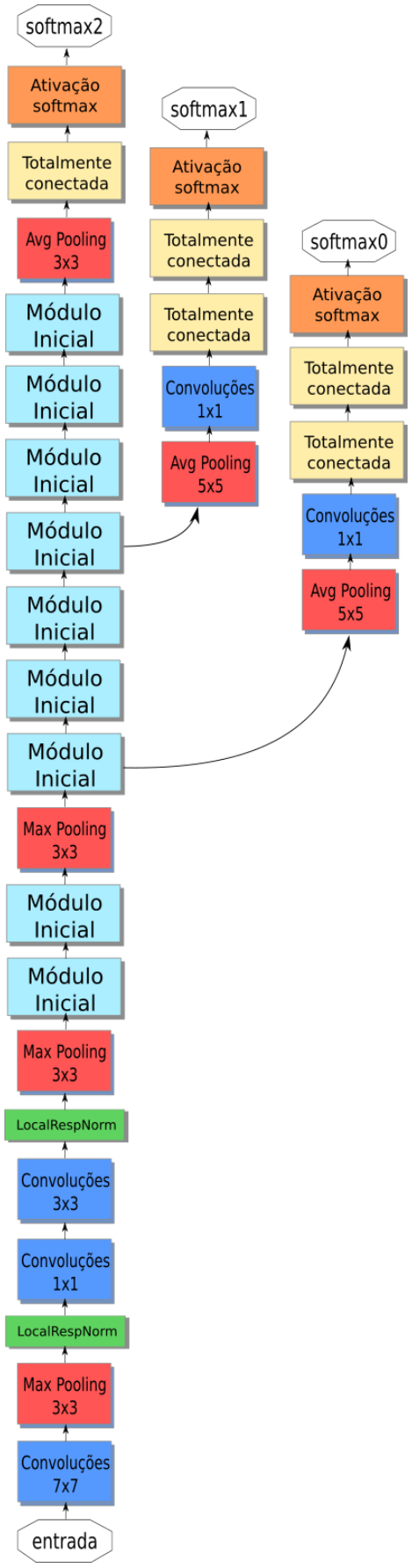


Figura 3.4 – Representação da arquitetura GoogLeNet.
Imagem baseada em (Szegedy et al. 2015).

3.1.3.2. Metodologia de Treinamento

Para o desafio ImageNet (Russakovsky et al. 2015), os autores treinaram a GoogLeNet (Szegedy et al. 2015) usando um algoritmo assíncrono estocástico gradiente descendente. Utilizou-se um momento de 0,9, juntamente com um aprendizado decrescente de 4% em cada intervalo de 8 épocas. Embora tenham sido aplicados alguns métodos de amostragem de imagens, os autores explicam que não havia um pipeline unificado para todos os modelos gerados para o desafio ImageNet (Russakovsky et al. 2015).

O *framework* de aprendizado profundo Caffe (Jia et al. 2014) possui uma distribuição da arquitetura GoogLeNet (Szegedy et al. 2015) e um modelo ImageNet. O modelo presente nessa distribuição foi treinado com algumas das técnicas de *data augmentation*, incluindo o recorte de diferentes regiões, e usou uma política de redução da taxa de aprendizado polinomial, em vez da política de etapa, porque permitia um treinamento cerca de quatro vezes mais rápido. Dados os recursos computacionais à nossa disposição, seria impossível treinar esta rede a partir do zero dentro de um tempo viável, usando as configurações originais. Além disso, dado que estes pesos são finalizados mais tarde para o nosso problema, espera-se que isto não cause impacto no desempenho. Por isso optamos por usar o modelo presente no *framework* Caffe (Jia et al. 2014) como nosso problema-fonte e adaptá-lo em dois níveis ao nosso problema-alvo (detecção de pornografia infantil).

3.2. DESCRITORES LOCAIS

Algumas abordagens existentes para a detecção de pornografia infantil usam métodos que utilizam o conceito de descritores locais de baixo nível. Estes métodos envolvem funções para a detecção de pontos de interesse na imagem, e para a descrição dos dados brutos nesse ponto e no seu entorno. No entanto, essas descrições de baixo nível ainda estão semanticamente longe de conceitos abstratos. Então, para se alcançar um desempenho satisfatório, faz-se necessária a realização de um passo intermediário antes da tomada de decisão final. Por esta razão, uma representação de nível médio é normalmente usada. Nesta seção, abordamos o descritor de características locais SURF (Bay et al. 2008), amplamente utilizado na literatura, bem como o Bossa-Nova (Avila et al. 2013), método de representação de médio nível que, juntamente com o SURF, são utilizados neste trabalho para comparação com a metodologia proposta.

3.2.1. *Speeded Up Robust Features* (SURF)

Visando projetar um detector de ponto de interesse, e descritor de imagem, rápido e eficiente – quando comparado ao SIFT (Lowe 2004) –, que fosse invariante à escala e rotação, (Bay et al. 2008) desenvolveram o SURF. Para detecção de pontos de interesse aplica-se a matriz Hessiana usando seu determinante como medida para escolher tanto a localização como a escala. O descritor é gerado empregando filtros de Haar no entorno do ponto de interesse e descrevendo a distribuição de respostas.

Apesar de usar conceitos semelhantes ao SIFT (Lowe 2004), o SURF reduziu a complexidade, conseguindo um método muito mais rápido. Com o objetivo de acelerar a técnica, são empregadas imagens integrais e filtros caixa, permitindo uma computação mais eficiente de operadores diferenciais aproximados.

3.2.2. Bossa-Nova

A abordagem de representação de médio nível mais utilizada na literatura, proposta por (Sivic & Zisserman 2003), é conhecida como Sacolas de Palavras Visuais (BoVW, do inglês *Bag of Visual Words*). Tradicionalmente, na representação BoVW, os descritores locais da imagem são associados ao elemento mais próximo do dicionário visual (*hard coding*) e, em seguida, a média desses descritores locais codificados é calculada, compactando toda a informação em um único vetor (*average pooling*).

Várias extensões da representação BoVW foram propostas na literatura. Entre elas, ressalta-se a representação Bossa-Nova (Avila et al. 2013). Em linhas gerais, esta representação segue o formalismo do modelo BoVW (*coding* & *pooling*), oferecendo um aprimoramento na etapa de *pooling*, a fim de preservar de uma maneira mais rica a informação obtida durante a etapa de *coding*. Assim, em vez de compactar toda a informação¸ relacionada a uma palavra visual em um único valor escalar, a etapa de *pooling* resulta em uma distribuição de distâncias. Para isto, é usada uma estimação não-paramétrica da distribuição dos descritores, calculando um histograma de distâncias entre os descritores encontrados na imagem e cada palavra visual do dicionário.

4. METODOLOGIA

Neste capítulo, são apresentados os detalhes da proposta para explorar o uso de técnicas de aprendizado profundo (*deep learning*) no problema de detecção automática de imagens de pornografia infantil.

A técnica de redes neurais convolucionais (*Convolutional Neural Networks* – CNN) foi escolhida pelos excelentes resultados apresentados nas tarefas de classificação de imagens e vídeos. Essa classe de técnicas foi utilizada nas abordagens campeãs do desafio de classificação de imagens ImageNet nos anos de 2014 e 2015.

Na Seção 4.1 são detalhados a arquitetura CNN adotada bem como o processo de treinamento dessa rede. A Seção 4.2 apresenta nosso roteiro para extração das informações das imagens a serem classificadas

4.1. ARQUITETURA GOOGLENET

Ao utilizar CNN, o primeiro passo consiste em escolher uma arquitetura apropriada para se trabalhar. A GoogLeNet, proposta por (Szegedy et al. 2015), foi a arquitetura de rede neural convolucional escolhida para os nossos experimentos. A escolha deste modelo de CNN para extração de características se deve ao fato de a GoogLeNet ter sido a campeã do desafio de classificação de imagens ImageNet 2014 (Russakovsky et al. 2015), atingindo a taxa de erro de 6,67% na competição. O conjunto de dados de treinamento da ImageNet possui cerca de 1,2 milhões imagens, com 1.000 classes dos mais variados assuntos, como objetos, pessoas, plantas e animais. Por esta razão, acredita-se que a GoogLeNet tenha a habilidade de aprender a extrair as características visuais mais significativas das imagens de entrada, e que um modelo treinado com este conjunto de dados manterá uma condição mais otimizada para extração de características de imagens. Por outro lado, a rede original foi proposta para um problema diferente do nosso dado que detectar pornografia e, na consequência pornografia infantil é um problema bastante diferente daquele de predizer a classe de um objeto em uma imagem.

A arquitetura GoogLeNet, com seus pesos, é uma solução acabada e está disponível *on-line*. Neste trabalho foi utilizada a distribuição de aprendizagem profunda presente no *framework* Caffe[4] (Jia et al. 2014).

Para a extração de características, foi escolhida a saída da última camada antes da classificação final, que tem uma dimensionalidade de 1.024. A camada de classificação do banco de filtros foi ajustada de 1.000 (número de classes na classificação ImageNet) para 2, a fim de adaptar os pesos para o nosso problema particular de detecção de duas classes (Pornografia Geral vs. Não Pornografia e, depois, para Pornografia Infantil vs. Não Pornografia Infantil). Todos os pesos de rede foram ajustados via *backpropagation*.

4.2. EXTRAÇÃO DE INFORMAÇÕES

Foram avaliadas duas abordagens de adaptação da arquitetura inicial para nosso problema. Assim, para extração de informações das imagens, foram avaliados dois modelos da arquitetura escolhida, aqui denominados "*deep features*" *e* "*fine tuning*", respectivamente. Estes modelos diferem, basicamente, na rede convolucional treinada, utilizada para extração das características de baixo nível.

4.2.1. *Deep Features*

No pipeline proposto para o modelo "*deep features*", a rede foi treinada a partir da configuração original para classificação de objetos no problema de detecção de pornografia geral (Perez 2016). A partir desse treinamento, a rede foi usada para extrair características das imagens a serem classificadas. Nesse ponto, a rede sabe o que é pornografia de forma direta, mas não tem um refinamento do que é pornografia infantil.

Além disso, como mostra a Figura 4.1, as imagens passam por um pré-processamento antes de serem encaminhadas à rede convolucional para extração de suas características. Por fim, a descrição das imagens é submetida a um classificador, para tomada de decisão final.

[4] http://caffe.berkeleyvision.org

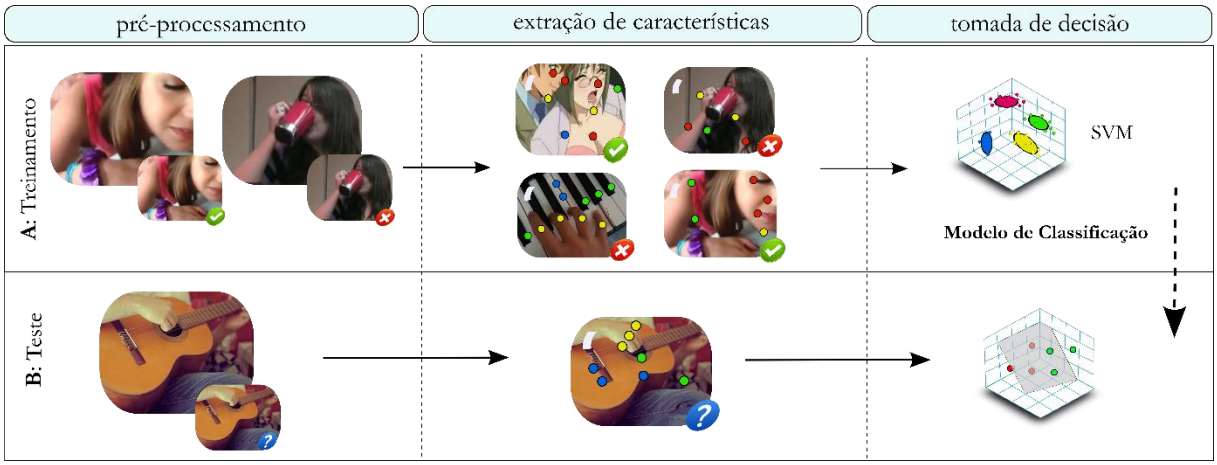


Figura 4.1 – Pipeline do modelo "*deep features*".

O pré-processamento é necessário para deixar a imagem com o formato necessário para arquitetura de rede convolucional escolhida. Para isso, cada imagem é redimensionada, mantendo a relação de aspecto, de modo que sua menor dimensão fique do mesmo tamanho da entrada da rede. Então é feito um recorte centralizado, resultando em uma imagem de 224x224 pixels.

Como falado anteriormente, a extração das características de baixo nível é realizada utilizando uma rede convolucional treinada especificamente com imagens de pornografia comum, gerada por (Perez 2016). Neste processo, chamado *transfer learning*, os pesos obtidos no treinamento de pornografia geral são aplicados ao nosso problema, utilizando-os na extração das características de baixo nível das novas imagens a serem processadas. Em seguida, as características das imagens são dadas como entrada para um algoritmo de aprendizagem de máquina (classificador de padrões) para a construção do modelo de classificação, e posterior tomada de decisão.

Dado que é sabido que o processo de treinamento de uma rede neural convolucional profunda pode requerer muitos dados, buscou-se, também, utilizar um processo de aumentação dos dados. Assim, adicionalmente, nesse caso, um novo conjunto de imagens de treinamento, obtido a partir do uso da técnica *data augmentation*, e um novo pipeline, mostrado na Figura 4.2, é seguido. Nesta técnica, são geradas seis versões de cada imagem do conjunto original de imagens de **treinamento** (V0), conforme ilustra Figura 4.3. A primeira versão (V1) é a imagem resultante do pré-processamento. A segunda versão (V2) é obtida com um recorte centralizado, na imagem original, de 640x640 e posterior redimensionamento para 224x224.

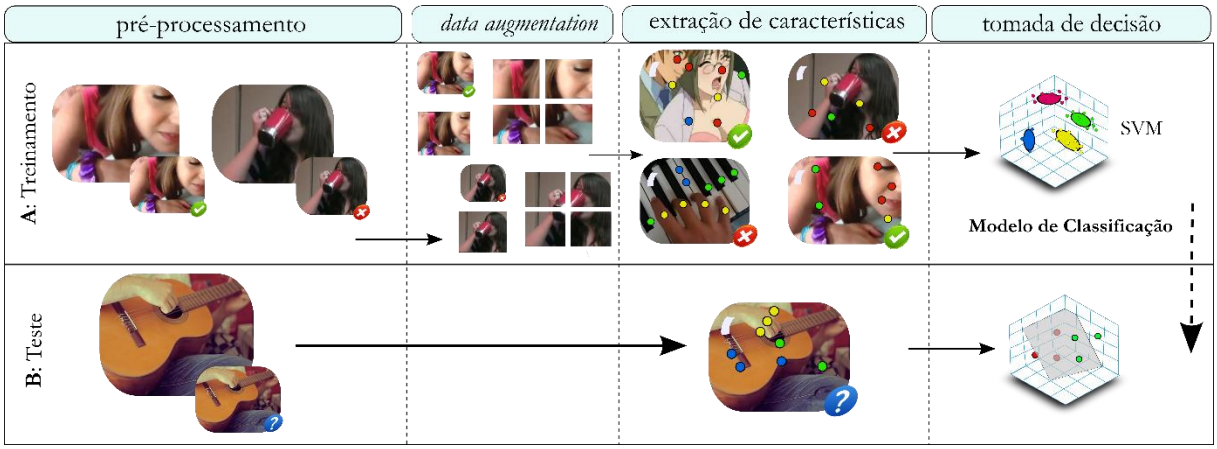


Figura 4.2 – Pipeline com *data augmentation.*

As demais versões são obtidas a partir do recorte de 480x480, na imagem original, tendo o centro da imagem original como canto inferior esquerdo (V3), ou canto inferior direito (V4), ou canto superior direito (V5), ou canto superior esquerdo (V6), e posterior redimensionamento para 224x224. Em todos os redimensionamentos, é mantida a relação de aspecto da imagem original. A quantidade de versões geradas foi considerada suficiente para os experimentos, razão pela qual não foram realizadas outras transformações, como rotação ou espelhamento, nas imagens originais. A técnica *data augmentation* busca aumentar a quantidade de pontos de interesse relevantes no conjunto de treinamento visando melhorar o modelo de classificação.

Figura 4.3 – A imagem original (V0) e as seis versões da imagem utilizadas no *data augmentation.*

4.2.2. *Fine Tuning*

Como discutido anteriormente, o refinamento dos pesos a partir da arquitetura original de classificação de objetos para o problema de classificação de pornografia geral é muito importante para que a rede passe a destacar características específicas de situações de pornografia e não de classes de objetos físicos de forma geral. No entanto, o grande objetivo aqui é a classificação de pornografia infantil. Assim, apresenta-se um segundo processo de refinamento da rede de modo que ela se especialize em problemas de pornografia infantil.

Nesse caso, a rede começa com características de problema geral de classificação de objetos em imagens e é refinada (primeiro estágio de refinamento) para o problema de pornografia em geral. A seguir, essa rede adaptada é novamente refinada (segundo estágio de refinamento) para o problema de pornografia infantil.

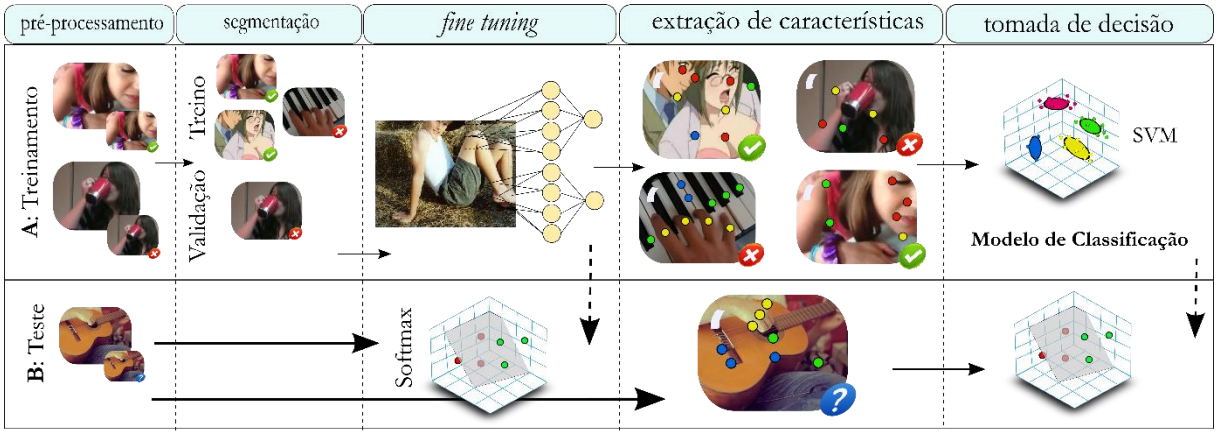


Figura 4.4 – Pipeline do modelo "*fine tuning*".

O pipeline proposto para o modelo "*fine tuning*", apresentado na Figura 4.4, consiste de duas fases distintas: treinamento e teste. Na fase de treinamento, após o pré-processamento, o conjunto de imagens de treinamento é dividido em treino e validação, e é utilizado para treinar um novo modelo da rede convolucional (*fine tuning*), que será utilizado para extração de características das imagens. Por fim, a descrição das imagens é submetida a um classificador, para a construção do modelo de classificação. Na fase de teste, após o pré-processamento, as imagens do conjunto de teste são submetidas à classificação pela camada softmax da nova rede convolucional gerada. Adicionalmente, a nova rede convolucional obtida no treinamento é utilizada na extração de características das imagens de teste, que são submetidas ao classificador, para tomada de decisão. Assim como no modelo "*deep features*",

o pré-processamento é necessário para deixar a imagem com o formato necessário para arquitetura de rede convolucional escolhida.

O conjunto de imagens de treinamento é então divido em dois grupos: treino e validação. O grupo treino é composto por 80% das imagens do conjunto de treinamento, e o grupo validação com os 20% restantes. Em ambos, a proporcionalidade entre imagens PI e não-PI do conjunto de treinamento original é mantida. Essa segmentação é realizada para que os pesos obtidos no treinamento da nova rede convolucional (segundo estágio de refinamento) sejam avaliadas no conjunto de validação, não sofrendo contaminação das imagens do conjunto de teste.

Esses grupos de imagens são então utilizados para treinar um novo modelo de rede convolucional. Durante o processo de aprendizagem da nova rede convolucional, os pesos obtidos da análise das imagens do grupo treino são validados no grupo validação, obtendo-se, assim, resultados parciais sobre a efetividade dos pesos no processo de classificação. A rede convolucional cujos pesos obtiveram o melhor resultado no processo de validação é então selecionada para utilização na extração das características de baixo nível.

As características extraídas das imagens são submetidas a um algoritmo de aprendizagem de máquina para a construção do modelo de classificação, e posterior tomada de decisão.

5. EXPERIMENTOS E RESULTADOS

Neste capítulo são apresentados os experimentos realizados para avaliar a metodologia proposta. Na Seção 5.1 são apresentadas as definições iniciais do experimento, como o conjunto de imagens, as métricas de avaliação, e detalhes do treinamento e de outros métodos existentes na literatura. Na Seção 5.2 são apresentados os experimentos e os resultados obtidos, bem como uma comparação desses resultados com os disponíveis na literatura.

5.1. DEFINIÇÕES INICIAIS

Nas próximas subseções são apresentadas as definições iniciais dos experimentos propostos para avaliar a metodologia proposta.

5.1.1. Conjunto de imagens

As imagens utilizadas nos experimentos foram obtidas junto ao Banco de Imagens de Pornografia Infantil (CP *Database*) do Setor Técnico-Científico da Superintendência de Polícia Federal no Paraná (SETEC/PR), instituído pela Instrução de Serviço (IS) 08/2012-SETEC/SR/PF/PR, de 26/09/2012.

Todas as imagens armazenadas neste banco foram obtidas de pastas, armazenadas nos discos rígidos analisados pelos peritos criminais do SETEC/PR, que continham algum material de pornografia infantil (PI). No entanto, o conteúdo dessas pastas não passa por uma triagem visual exaustiva, razão pela qual este banco pode conter imagens que, apesar de manterem forte relação com este tipo de conteúdo como, por exemplo, crianças em trajes de banho, não configuram pornografia infantil (não-PI). Isso se deve ao fato de que nem toda imagem que é atraente para indivíduos com interesse sexual em crianças é necessariamente ilegal (Taylor et al. 2001).

Para o nosso experimento, foram selecionadas no CP *Database*, através da análise visual, 58.971 imagens, sendo 33.723 imagens PI, e 25.248 imagens não-PI. O conjunto de treinamento ficou com 39.584 imagens, e o conjunto de teste com 19.387 imagens, mantida a proporcionalidade entre imagens PI e não-PI do *dataset* (aproximadamente, 60/40%) nos conjuntos.

Na seleção dessas imagens, buscou-se garantir uma boa diversidade. Em relação ao tamanho, as imagens variam de 10 mil a 10 milhões de pixels. Quanto à qualidade visual, foram selecionadas tanto imagens caseiras, com pouca iluminação ou desfocadas, quanto imagens profissionais, produzidas por publicações eletrônicas especializadas na divulgação desse tipo de conteúdo (PI). De um mesmo "*Book* Fotográfico" foram selecionadas tanto imagens PI quanto não-PI. Já em relação ao conteúdo, as imagens apresentam uma variedade de ângulos e *close-ups*, não havendo qualquer tipo de preferência quanto à prevalência de rosto, genitália ou posição corporal. Da mesma forma, o ambiente retratado nas imagens selecionadas apresenta desde paisagens ao ar livre, como praia ou campo, até ambientes confinados como quartos ou carros.

Devido à vedação legal (Brasil 2016), as imagens de PI foram manipuladas exclusivamente por agente público no exercício de suas funções, e por obrigatoriedade de sigilo, não poderão ilustrar este trabalho.

5.1.2. Métricas de avaliação

As métricas escolhidas para avaliação dos resultados da metodologia proposta foram acurácia normalizada (ACC) e medida F_2 (F_2 *measure*). A acurácia normalizada é a média aritmética das porcentagens das classes, PI e não-PI, classificadas corretamente. Ao contrário do que ocorre na acurácia, na acurácia normalizada os melhores resultados são obtidos quando o classificador atinge um alto índice acertos nas duas classes. A acurácia normalizada é definida conforme a Equação 5.1.

$$ACC = \frac{True\ Positive(\%) + True\ Negative(\%)}{2} \qquad \text{(Eq. 5.1)}$$

A medida F_2 é a média harmônica ponderada entre precisão e revocação (*recall*), dando à revocação o dobro do peso da precisão (β = 2). O uso da medida F_2 é importante para se avaliar o impacto dos resultados de falso negativo. Nos casos de detecção de pornografia infantil, é preferível ter imagens não-PI classificadas como PI (falso positivo), ao inverso, imagens de PI classificadas como não-PI (falso negativo). A medida F_2 é definida conforme a Equação 5.2.

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall} \quad \text{(Eq. 5.2)}$$

onde β é a majoração de peso da revocação sobre a precisão.

5.1.3. Configuração do método proposto

Como mencionado anteriormente, o *dataset* inicialmente utilizado foi particionado em dois conjuntos, um para treinamento, com 39.584 imagens, e outro para teste, com 19.387 imagens, mantendo-se nesses conjuntos a proporção entre imagens PI e não-PI do *dataset* original (aproximadamente, 60/40).

A abordagem proposta neste trabalho para detecção automática de pornografia infantil é baseada na utilização de redes neurais convolucionais para a extração das características de baixo nível das imagens, e posterior aplicação de um classificador.

Inicialmente, as imagens devem passar por um pré-processamento, necessário para deixar a imagem com o formato necessário para arquitetura de rede convolucional escolhida, resultando, nosso caso, em imagens de 224x224 pixels.

Na primeira fase de processamento, aqui denominada "*deep features*", é realizada a extração das características de baixo nível utilizando um modelo de rede convolucional previamente treinado com imagens de pornografia comum. Então, a essas características é aplicado o classificador SVM, usando kernel RBF implementado pela biblioteca LIBSVM[5] (Chang & Lin 2011). Foi aplicada a busca em *grid* para encontrar o melhor parâmetro *C* do SVM durante o treinamento. Também foi avaliada a utilização do classificador *Random Forest*, implementado pela biblioteca FEST[6]. Para este classificador, foram utilizados diversos valores para o parâmetro *t* (número de árvores), compreendidos entre 25 e 800, sendo o melhor resultado obtido com 200 árvores. Adicionalmente, buscou-se melhorar os resultados deste protocolo com a utilização da técnica *data augmentation*, o que, frustrando expectativas, não aconteceu, pelos motivos expostos mais à frente.

[5] LibSVM: https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[6] FEST: http://lowrank.net/nikos/fest/

Na fase seguinte, aqui denominada "*fine tuning*", o conjunto de imagens de treinamento é segmentado em treino e validação, na proporção de 80/20% das imagens em cada segmento, e é utilizado para treinar um novo modelo da rede convolucional, a partir do modelo GoogLeNet ImageNet previamente treinado. Neste processo, o modelo de rede convolucional que obteve os melhores resultados na validação é selecionado e utilizado na extração das características de baixo nível. Como na fase anterior, às características é aplicado o classificador LIBSVM. Por não apresentarem melhora no resultado obtido com o LIBSVM nos experimentos realizados na fase "*deep learning*", o classificador *Random Forest* e a técnica *data augmentation* não foram consideradas nesta fase. Destaca-se, porém, que outros experimentos deverão ser realizados antes de se concluir que o classificador *Random Forest* não é o mais adequado para o novo problema. No entanto, outro processo de classificação foi testado, passando-se as imagens do conjunto de teste para classificação pela camada softmax da rede convolucional obtida nesta fase, obtendo-se o resultado idêntico que o obtido com o LIBSVM. Adicionalmente, foram incluídas no conjunto de treinamento 20.100 imagens não-PI, obtidas do conjunto de teste do *dataset* MSCOCO[7] 2014 (do inglês, Microsoft *Common Objects in Context*) (Lin et al. 2014), na tentativa de melhorar o acerto da classe negativa (e positiva, por consequência), com o incremento de imagens sem qualquer relação com pornografia infantil.

O *framework* Caffe[8] (Jia et al. 2014) foi utilizado como plataforma para realização de extração de características e treinamento do modelo da rede convolucional. Neste *framework*, o modelo GoogLeNet ImageNet foi treinado mais rápido quando utilizada a diretiva *polynomial learning rate decay*, razão pela qual esta diretiva foi adotada nos experimentos. Para cada tipo ou fonte de informação, é possível alterar os valores de *base learning rate*, *weight decay*, *polynomial power* e o número máximo de épocas (*epochs*). Os valores desses hiper-parâmetros utilizados no modelo proposto, apresentados na Tabela 5.1, foram obtidos por experimentação durante o treinamento.

Tabela 5.1 – Hiper-parâmetros de aprendizado.

Taxa de Aprendizado (*Learning Rate*)	**Decaimento (*Weight Decay*)**	**Potência (*Power*)**	**Número Max. Épocas (*Max Epochs*)**
0,000009	0,005	0,5	200

[7] MSCOCO: http://mscoco.org/
[8] Caffe: http://caffe.berkeleyvision.org/

5.1.4. Comparação com métodos existentes

A comparação com outros métodos propostos na literatura é essencial para melhor interpretar os resultados obtidos com o método proposto. Foram avaliados métodos de detecção de tons de pele e de sacola de palavras visuais (BoVW). Esses métodos foram escolhidos por serem muito utilizados em problemas de detecção de pornografia tradicional e, eventualmente, de detecção de pornografia infantil.

Técnicas de detecção de pele humana são amplamente aplicadas para classificação de pornografia infantil (Polastro & Eleuterio 2010; Islam et al. 2011). Diversas soluções para detecção de pele têm sido propostas na literatura (Kelly et al. 2007), sendo que a forma mais comum e intuitiva utiliza características de cor (Kakumanu et al. 2007). Por ser uma técnica bastante utilizada e de fácil implementação, o detector de pele de (Kovac et al. 2003) foi reimplementado neste trabalho para fins de comparação. Neste caso, as imagens de treinamento foram utilizadas para se definir o *threshold* que melhor resolvia o problema de detecção de pornografia infantil.

Outra abordagem utilizada no problema de detecção automática de pornografia infantil é baseada no modelo Sacolas de Palavras Visuais (BoVW, do inglês *Bag of Visual Words*). Neste trabalho, utilizamos uma solução baseada em uma representação multi-nível, consistindo em um mapeamento suave de características de baixo nível semântico para características de maior nível conceitual associado. Para essa representação, lançamos mão do conceito de descritores locais, dicionários visuais com preservação de informação de distância e classificação de padrões.

Para a etapa de baixo nível, primeiramente, as imagens são redimensionadas para, no máximo, 100k pixels, economizando o tempo de processamento das etapas posteriores (Akata et al. 2014). Em seguida, descritores SURF (Bay et al. 2008) são extraídos usando uma amostragem densa com cinco escalas diferentes. Precisamente, usamos regiões de tamanhos 24×24, 32×32, 48×48, 68×68 e 96×96 pixels, com espaçamento de 4, 6, 8, 11 e 16 pixels, respectivamente. Além disso, a dimensionalidade do descritor SURF é reduzida, de 64 para 32 dimensões, utilizando Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*). Os códigos do descritor SURF e do PCA foram obtidos a partir do repositório OpenCV[9], uma das bibliotecas mais populares de Visão Computacional. Para a etapa de médio nível, são extraídos vetores Bossa-Nova (Avila et al. 2013), mantendo os

[9] OpenCV: http://opencv.org

valores padrão[10], e vetores BoVW (Sivic & Zisserman 2003) (*hard coding e average pooling*), para fins de comparação. O tamanho dos dicionários é variado em {512, 1024, 2048} palavras visuais. Por fim, na etapa de alto nível, classificadores SVMs são aplicados, utilizando um kernel χ^2 implementado pela biblioteca PMSVM[11] (*Power Mean* SVM) (Wu 2012).

5.1.5. Ferramentas forenses

Apesar de existirem várias ferramentas forenses para combater a pornografia infantil, a maioria não está disponível para o público em geral, e poucas soluções analisam o conteúdo visual das imagens para classifica-las. Assim, para avaliar o desempenho da solução proposta, foram selecionadas as seguintes ferramentas: NuDetective (Polastro & Eleuterio 2010), Localizador de Evidencias Digitais (LED), MediaDetective[12] e Snitch Plus[13]. O NuDetective pode ser adquirido gratuitamente por agencias de aplicação da lei ou para fins de pesquisa, enquanto que a ferramenta LED está disponível apenas para os peritos criminais da Polícia Federal do Brasil.

O MediaDetective e o Snitch Plus são soluções comerciais para filtragem de conteúdo pornográfico. Todos estes sistemas analisam o conteúdo das imagens por meio de técnicas de detecção de pele, principalmente. Além disso, para o MediaDetective e o Snitch Plus, as imagens são classificadas de acordo com o valor de probabilidade. Ou seja, a imagem tem conteúdo pornográfico infantil se a probabilidade for igual ou maior que 50%.

De modo similar, para o LED, as imagens são classificadas de acordo com o valor do Detector de Imagens Explícitas (DIE), que varia de 0 a 1000. Para os nossos experimentos, a imagem tem conteúdo pornográfico infantil se o valor do DIE for igual ou maior que 500. O NuDetective, por outro lado, atribui valores binários para a imagem: positivo (pornografia infantil) ou negativo (não pornografia infantil). Por fim, o MediaDetective e o Snitch Plus têm quatro modos de execução pré-definidos, que diferem principalmente quanto ao rigor do detector de pele. Nos experimentos, optou-se pelo modo mais rigoroso. Em relação ao NuDetective e ao LED, empregou-se as configurações padrão.

[10] Bossa-Nova: https://sites.google.com/site/bossanovasite
[11] PmSVM: https://sites.google.com/site/wujx2001/home/power-mean-svm
[12] MediaDetective: http://www.mediadetective.com
[13] Snitch Plus: http://www.hyperdynesoftware.com

5.2. RESULTADOS

Nesta seção são apresentados e discutidos os resultados obtidos com os experimentos realizados. Na Subseção 5.2.1 é avaliada a abordagem proposta, e na Subseção 5.2.2 o resultado desta abordagem é comparado com os resultados de outros métodos da literatura.

5.2.1. Abordagem proposta

A acurácia (ACC) e a medida F_2 (F_2) da classificação das imagens realizada com a abordagem proposta é apresentada na Tabela 5.2. Nessa tabela, podemos ver que o modelo baseado no aprendizado prévio com imagens de pornografia comum (*deep features*) obteve resultados significativos, com 80,37% de acurácia e 83,42% de F_2. Porém, com o aperfeiçoamento do modelo originalmente treinado para pornografia geral e melhor adaptação ao problema de pornografia infantil com pesos propriamente escolhidos para esse problema (*fine tuning*), houve um incremento expressivo nos resultados, elevando-os a 86,06% de acurácia e 87,73% de F_2, o que corresponde a uma melhora de mais de 5 pontos percentuais nos resultados. Já com a adição de imagens não-PI do *dataset* MSCOCO 2014 (aumentação dos dados), houve apenas uma pequena melhoria na acurácia, mas uma redução em F_2.

Da análise da Tabela 5.2, observa-se, na fase *deep features*, a superioridade do classificador LIBSVM sobre o *Random Forest,* para o problema de detecção automática de pornografia infantil, com um ganho de cinco pontos percentuais na acurácia, e de três pontos percentuais em F_2. Vemos também que a utilização da técnica de aumentação dos dados (*data augmentation*), nesse problema em específico e para as imagens consideradas, não surtiu efeito na melhora dos resultados obtidos com a utilização do LIBSVM sobre as características extraídas do conjunto de teste original. Isto se deve ao fato de que, nas imagens de PI, os pontos de interesse estão concentrados em poucas regiões das imagens. Assim, ao realizar as transformações para aumentação de dados, são geradas versões não-PI de uma imagem positiva para PI, o que acaba por dificultar a classificação.

A fase de refinamento do método (*fine tuning*), por sua vez, suplantou os excelentes resultados obtidos na fase *deep features*, já na utilização da camada softmax da CNN como classificador. Como esta camada não fornece a matriz de confusão, não possível calcular o valor de F_2 neste ponto. Com a aplicação do classificador LIBSVM, veio a confirmação da boa acurácia obtida, e o melhor resultado para F_2. Com o incremento das imagens não-PI do

dataset MSCOCO no conjunto de treinamento, a taxa de verdadeiros negativos (TVN) alcançou a marca dos 97%, mas, em contrapartida, reduziu a taxa de verdadeiros positivos (TVP), o que levou a um decréscimo em F_2, apesar do pequeno ganho na acurácia.

Tabela 5.2 – Resultados da abordagem proposta.

Solução Proposta		**ACC (%)**	**F_2 (%)**
Deep features	*Random Forest*	75,35	80,12
	Data augmentation	76,14	80,62
	LIBSVM	**80,37**	**83,42**
Fine tuning	Softmax	86,06	- -
	LIBSVM	86,08	**87,73**
	MSCOCO (aumentação dos dados)	**86,47**	87,12

5.2.2. Comparação com outras abordagens

A Tabela 5.3 apresenta os resultados de classificação das imagens, obtidos utilizando-se outras abordagens comumente utilizadas para detecção de pornografia infantil, como detecção de tons de pele e modelo BoVW, bem como os resultados obtidos pelas ferramentas forenses avaliadas.

Como se pode observar, a abordagem baseada no modelo BoVW supera a solução do detector de pele e as ferramentas forenses analisadas. No que concerne à F_2, como as soluções baseadas em detecção de pele reportam baixo número de falsos negativos — e alto número de falsos positivos —, os valores de F_2 são consequentemente mais altos. No entanto, para detecção de conteúdo pornográfico infantil, é essencial obter bons resultados em relação às duas métricas, ACC e F_2. Para o modelo BoVW, o melhor resultado foi obtido com 2.048 palavras visuais.

Ao se comparar esses resultados com os obtidos pela abordagem proposta, verifica-se que as redes neurais convolucionais estão em um patamar muito superior quando se trata do problema de classificação de imagens de pornografia infantil. Quando comparada com as técnicas de detecção de tons de pele, o *fine tuning* teve um ganho na acurácia em torno de 30 pontos percentuais, e superior a 20 pontos percentuais quando comparada com o modelo BoVW. O alto valor de F_2 corrobora a superioridade da abordagem proposta, pois demonstra a sua competência tanto na classificação de imagens PI, quando na de imagens não-PI.

Assim, vê-se que dentre as soluções avaliadas, baseadas em detecção de tons de pele e em BoVW, a solução proposta fornece classificadores mais eficazes, tanto em termos de ACC (86,5%) como de F2 (87,1%).

Tabela 5.3 – Resultados com outras soluções.

Solução		**ACC (%)**	**F_2 (%)**
Tons de pele	Kovac	53,05	82,41
Ferramentas forenses	LED	54,21	84,38
	MediaDetective	56,61	70,01
	NuDetective	56,68	73,84
	Snitch Plus	56,79	67,83
BoVW	Bossa-Nova	**68,25**	**71,10**
Abordagem proposta	*Deep features*	**80,37**	**83,42**
	Fine tuning	**86,47**	**87,12**

6. CONCLUSÕES

As abordagens apresentadas neste trabalho, para a detecção de pornografia infantil em imagens, confirmam a superioridade da aprendizagem profunda (*deep learning*) em relação às ferramentas forenses, e sobre as soluções tipicamente encontradas na literatura para a detecção automática de pornografia tradicional e, eventualmente, de detecção de pornografia infantil.

A avaliação da solução baseada no modelo sacolas de palavras visuais (BoVW) (Vitorino et al. 2016) já produziu um resultado competitivo em relação a métodos de detecção de tons de pele, atingindo um acréscimo na acurácia normalizada de mais de 10 pontos percentuais, um ganho considerável para um problema com um conceito tão subjetivo como este, e onde todas as imagens utilizadas, positivas ou negativas para pornografia infantil, foram obtidas de discos rígidos de casos reais, examinados por peritos criminais da Polícia Federal do Brasil (PF).

Ao se analisar os resultados obtidos pela abordagem proposta, verifica-se que as redes neurais convolucionais (CNNs) estão em um patamar muito superior quando se trata do problema de classificação de imagens de pornografia infantil. O modelo baseado no aprendizado prévio com imagens de pornografia comum (*deep features*), que utiliza apenas um nível de transferência de conhecimento de um problema-fonte (classificação de imagens naturais para um problema-alvo, detecção de pornografia geral), obteve resultados significativos, com 80,37% de acurácia, superando o modelo BoVW em mais de 10 pontos percentuais. Porém, com o aperfeiçoamento do modelo originalmente treinado para pornografia geral e melhor adaptação ao problema de pornografia infantil com pesos propriamente escolhidos para esse problema (*fine tuning*) em uma segunda etapa de transferência de conhecimento, houve um incremento expressivo nos resultados, elevando a acurácia de classificação a 86,06%.

O discreto aumento de 0,4 pontos percentuais na acurácia, com a adição de imagens não-PI do *dataset* MSCOCO 2014 (aumentação dos dados), onde a taxa de verdadeiros negativos (TVN) atingiu a expressiva marca de 97%, mostra que essa técnica pode ser melhor explorada para obtenção de resultados mais significativos, em especial, com o aumento das imagens positivas para pornografia infantil. No entanto, apesar da abordagem proposta ter apresentado resultados promissores, a detecção automática de pornografia infantil continua sendo um problema em aberto. Faz-se necessário, ainda, avaliar se a arquitetura CNN utilizada pode ser

melhorada ainda mais para esse problema em específico. Ademais, com os resultados obtidos por (Perez 2016) na utilização de CNNs para detecção automática de pornografia tradicional em vídeos, vislumbra-se a extensão deste trabalho para a detecção automática de pornografia infantil em vídeos.

Por fim, diante da inegável capacidade das redes neurais convolucionais na classificação de imagens, e dos ótimos resultados obtidos neste trabalho, entende-se que as CNNs podem ser avaliadas na classificação de imagens de pornografia infantil não apenas em duas classes, positiva ou negativa, mas em número suficiente para atender a alguma escala específica, como a proposta pelo Projeto COPINE (Quayle 2008). A escala COPINE estabelece 10 níveis de classificação de material de pornografia infantil, de acordo com o grau de violência, ou vitimização, nele representado. Em (Smid et al. 2014), os autores mostraram a importância da utilização dessa escala na priorização de investigações relativas à pornografia infantil, e no resgate precoce de vítimas desse tipo de violência.

REFERÊNCIAS BIBLIOGRÁFICAS

Akata, Z. et al., 2014. Good Practice in Large-Scale Learning for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3), pp.507–520. Available at: https://hal.inria.fr/hal-00835810/document.

Avila, S. et al., 2013. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5), pp.453–465. Available at: http://dx.doi.org/10.1016/j.cviu.2012.09.007.

Bay, H. et al., 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3), pp.346–359.

Bouirouga, H. et al., 2012. Skin Detection in Pornographic Videos Using Threshold Technique. *Journal of Theoretical and Applied Information Technology*, 35(1), pp.7–19.

Brasil, 2016. *Estatuto da criança e do adolescente [recurso eletrônico] : Lei n. 8.069, de 13 de julho de 1990, e legislação correlata.* 14ª ed., Brasília: Câmara dos Deputados, Edições Câmara. Available at: http://bd.camara.gov.br/bd/bitstream/handle/bdcamara/18403/estatuto_crianca_adolescente_14ed.pdf?sequence=40.

Carvalho, I.A. de, 2012. *Classificação de Imagens de Pornografia e Pornografia Infantil Utilizando Recuperação de Imagens Baseada em Conteúdo*. Universidade de Brasília.

Chang, C. & Lin, C., 2011. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, pp.1–39.

Deselaers, T., Pimenidis, L. & Ney, H., 2008. Bag-of-visual-words models for adult image classification and filtering. In *19th International Conference on Pattern Recognition*. IEEE, pp. 1–4. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4761366.

Digiácomo, M.J. & Digiácomo, I. de A., 2013. *Estatuto da Criança e do Adolescente Anotado e Interpretado* 6ª ed., Curitiba: Ministério Público do Estado do Paraná.

Fleck, M.M., Forsyth, D.A. & Bregler, C., 1996. Finding Naked People. *Journal of Chemical Information and Modeling*, 1065, pp.593–602.

Forsyth, D.A. & Fleck, M.M., 1999. Automatic Detection of Human Nudes. *International Journal of Computer Vision (IJCV)*, 32(1), pp.63–77.

Hochreiter, S. et al., 2000. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In *IEEE Press Field Guide to Dynamical Recurrent Networks*. pp. 237–243.

Huang, Y. & Kong, A.W.K., 2016. Using a CNN Ensemble for Detecting Pornographic and Upskirt Images, pp. 1-7, IEEE Intl. Conference on Biometrics: Theory, Applications, and Systems.

Islam, M., Watters, P.A. & Yearwood, J., 2011. Real-time detection of children's skin on social networking sites using Markov random field modelling. *Information Security Technical Report*, 16(2), pp.51–58. Available at: http://dx.doi.org/10.1016/j.istr.2011.09.004.

Jansohn, C., Ulges, A. & Breuel, T.M., 2009. Detecting Pornographic Video Content by Combining Image Features with Motion Information. *Proceedings of the Seventeen ACM International Conference on Multimedia - MM '09*, pp.601–604. Available at: http://portal.acm.org/citation.cfm?doid=1631272.1631366.

Jia, Y. et al., 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM Intl. Conference on Multimedia*. pp. 675–678.

Jones, M.J. & Rehg, J.M., 2002. Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision*, 46(1), pp.81–96. Available at: http://link.springer.com/10.1023/A:1013200319198.

Kakumanu, P., Makrogiannis, S. & Bourbakis, N., 2007. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3), pp.1106–1122.

Kawale, N. & Patil, S., 2014. An Approach To Maintain The Stroage Of Contentious Image In The Form Of Descriptor. In *IEEE International Conference on Computational Intelligence and Computing Research*.

Kelly, W., Donnellan, A. & Molloy, D., 2007. A Review of Skin Detection Techniques for Objectionable Images. *IT&T 2007 General Chair's Letter*, pp.40–49. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.4087&rep=rep1&type=pdf#page=53.

Kovac, J., Peer, P. & Solina, F., 2003. Human skin color clustering for face detection. *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, 2, pp.144–148 vol.2.

Krizhevsky, A., Sutskever, I. & Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. pp. 1097–1105.

Lee, S., Shim, W. & Kim, S., 2009. Hierarchical system for objectionable video detection. *IEEE Transactions on Consumer Electronics*, 55(2), pp.677–684.

Lin, T.Y. et al., 2014. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5), pp.740–755.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision (IJCV)*, 60(2), pp.91–110.

Microsoft, 2009. PhotoDNA. Available at: https://ncmedia.azureedge.net/ncmedia/2016/03/photoDNACloudServiceFS.pdf.

Moustafa, M., 2015. Applying deep learning to classify pornographic images and videos. *arXiv:1511.08899 [cs]*. Available at: http://arxiv.org/abs/1511.08899.

Oliveira, J.R.S. de & Silva, E.E. da, 2009. EspiaMule e Wyoming ToolKit: Ferramentas de Repressão à Exploração Sexual Infanto-Juvenil em Redes Peer-to-Peer. In *Proceedings of The Fourth International Conference on Forensic Computer Science*. pp. 108–113.

Perez, M.L., 2016. *Video pornography detection through deep learning techniques and motion information*. Universidade Estadual de Campinas.

Polastro, M.D.C.M. & Eleuterio, P.M.D.S., 2010. NuDetective: A Forensic Tool to Help Combat Child Pornography through Automatic Nudity Detection. *2010 Workshops on Database and Expert Systems Applications*, pp.349–353. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5591120 [Accessed September 26, 2014].

Quayle, E., 2008. The COPINE Project. *Irish Probation Journal*, 5, pp.65–83.

Russakovsky, O. et al., 2015. ImageNet Large Scale Visual Recognition Challenge. *Intl. Journal of Computer Vision (IJCV)*, 115(3), pp.211–252.

Sae-Bae, N. et al., 2014. Towards automatic detection of child pornography. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 5332–5336. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7026079.

Schulze, C. et al., 2014. Automatic Detection of CSA Media by Multi-modal Feature Fusion for Law Enforcement Support. *Proceedings of International Conference on Multimedia Retrieval - ICMR '14*, pp.353–360. Available at: http://dl.acm.org/citation.cfm?doid=2578726.2578772.

Short, M.B. et al., 2012. A review of Internet pornography use research: Methodology and content from the past 10 years. *Cyberpsychology, Behavior, and Social Networking*, 15(1), pp.13–23.

Sivic, J. & Zisserman, A., 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. *Proc. of ICCV2003*, pp.1470–1477.

Smid, W. et al., 2014. Prioritizing Child Pornography Notifications: Predicting Direct Victimization. *Sexual Abuse: A Journal of Research and Treatment*, 26(3), pp.1–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24492619.

Steel, C.M.S., 2012. The Mask-SIFT cascading classifier for pornography detection. *Internet Security (WorldCIS), 2012 World Congress on*, pp.139–142. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6280215.

Szegedy, C. et al., 2015. Going Deeper with Convolutions. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1–9.

Taylor, M. a X., Holland, G. & Quayle, E., 2001. Typology of paedophile picture collections. *The Police Journal*, 74, pp.97–107. Available at: http://www.researchgate.net/publication/237641206.

Ulges, A. & Stahl, A., 2011. Automatic detection of child pornography using color visual words. *2011 IEEE International Conference on Multimedia and Expo*, pp.1–6.

Vitorino, P., Avila, S. & Rocha, A., 2016. A Two-tier Image Representation Approach to Detecting Child Pornography. In *XII Workshop de Visão Computacional*. pp. 129–134.

Vrubel, A., 2011. Creation and Maintenance of MD5 Hash Libraries, and Their Application in Cases of Child Pornography. In *Proceedings of The Sixth International Conference on Forensic Computer Science*. ABEAT, pp. 137–141. Available at: http://www.icofcs.org/2011/papers-published-015.html.

Wang, X., Guo, R. & Kambhamettu, C., 2015. Deeply-learned feature for age estimation. *Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, pp.534–541.

Wu, J., 2012. Power mean SVM for large scale visual classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2344–2351.

Zeiler, M.D. & Fergus, R., 2014. Visualizing and understanding convolutional neural networks. In *Springer European Conference on Computer Vision (ECCV)*. pp. 818–833.

Zheng, H., Daoudi, M. & Jedynak, B., 2004. Blocking Adult Images Based on Statistical Skin Detection. *Electronic Letters on Computer Vision and Image Analysis*, 4(2), pp.1–14.