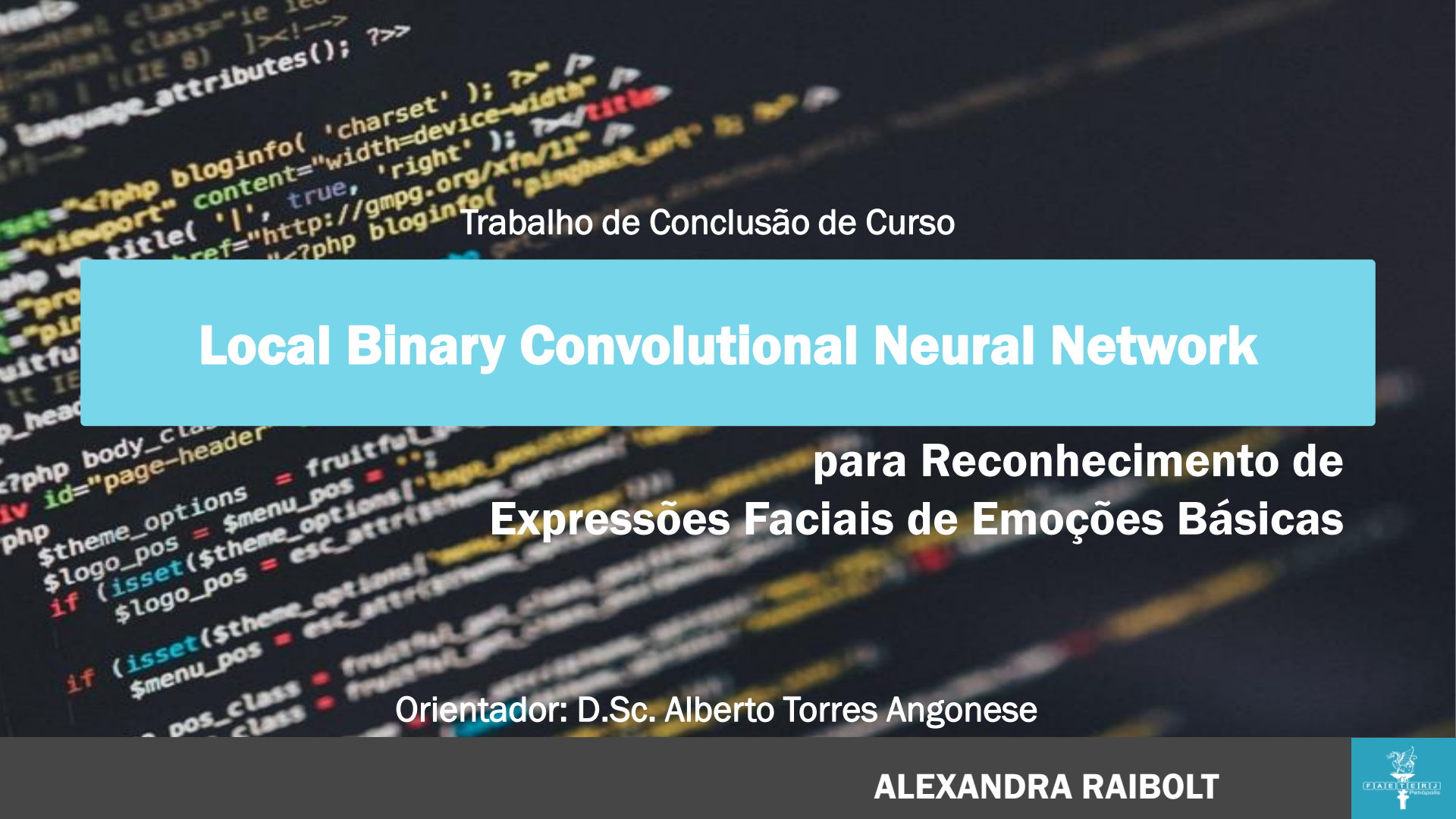


Trabalho de Conclusão de Curso

Local Binary Convolutional Neural Network

para Reconhecimento de Expressões Faciais de Emoções Básicas

Orientador: D.Sc. Alberto Torres Angonese

ALEXANDRA RAIBOLT

Sumário

- Introdução;
  - Motivação;
  - Justificativa;
  - Objetivos;

- Fundamentação Teórica;
  - Redes Neurais Convolucionais;
  - Padrão Binário Local;
  - Reformulação do LBP através de filtros convolucionais;

Sumário

- Implementação;
  - Base de imagens;
  - Pré-processamento;
    - Formato .pkl;
    - Codificação *One-Hot*;
  - Rede Neural Convolucional Binária Local;

- Experimentos e Resultados;

- Considerações Finais;
  - Trabalhos Futuros.

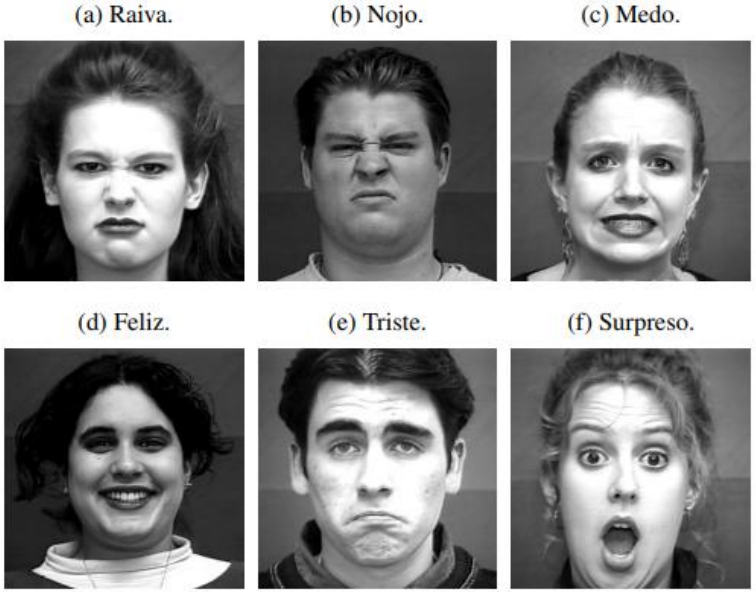
Introdução

- As expressões faciais são um meio de comunicação não verbal durante o processo de comunicação, que fornece informações de um indivíduo sobre suas intenções, desejos, objetivos, estado de espírito, etc.;

- Seis emoções básicas ou universais (Raiva, Nojo, Medo, Alegria, Tristeza, Surpresa.), foram foco de estudo de Ekman, e consideradas por ele como sendo as emoções comuns aos seres humanos, independente de fatores culturais.

Introdução

Figura 1: As seis expressões faciais de emoções básicas.



Fonte: Adaptado de [1, 2].

Introdução

- Nas relações interpessoais dos seres humanos realizar a tarefa de reconhecimento de objetos, reconhecimento facial, reconhecimento de expressões faciais torna-se uma tarefa fácil e corriqueira;

- Entretanto, computacionalmente tal tarefa apresenta um alto grau de complexidade.

Introdução

- Devido à grande necessidade encontrada na interação robô-humano por sistemas de reconhecimento automatizados, reconhecimento automatizado de objetos, facial, e de expressões faciais torna-se um importante desafio no campo da Visão Computacional;

- Pois dispõe de incontáveis aplicações do mundo real, como robótica Assistiva, gerenciamento de estoque, sistema de vigilância, controle automático de acesso, detecção de transtornos mentais, autenticação, etc.

Introdução

- O crescente estudo de Aprendizado de Máquina e Reconhecimento de Padrões nas últimas duas décadas para solucionar tarefas de Visão Computacional, conduziu pesquisas em busca de técnicas de extração de recursos de imagens;

- Dentre estas pesquisas, algumas técnicas apresentam bons resultados no processo de extração de características e padrões de classificação em imagens de faces para o reconhecimento e classificação de expressões faciais, como é o caso de: SVM, Redes Bayesianas e CNN.

Motivação

- As arquiteturas de CNN vêm conquistando espaço em desafios de reconhecimento e classificação de imagens a partir do ano de 2012 no desafio Imagenet;

- Seus resultados as tornam estado-da-arte para a resolução de problemas na área de Visão Computacional;

- Entretanto, um problema ainda enfrentado ao treinar tais arquiteturas, está relacionado ao poder computacional necessário, onde, torna-se um recurso caro, ou até mesmo indisponível;

- Em contrapartida, a arquitetura do LBCNN [3] demonstra um desempenho computacional eficiente, com sua complexidade computacional reduzida em comparação a outras arquiteturas de CNN.

Justificativa

- A partir de pesquisas realizadas em 2017 pela Federação Internacional de Robótica [4, 5] é possível observar que:
  - (1) Há uma discrepância na comparação do Brasil com outros países, como por exemplo aspectos socioeconômicos, ao fato de países desenvolvidos investirem no seu desenvolvimento tecnológico, apoiando financeiramente a educação e a pesquisa, enquanto no Brasil há déficits no apoio e investimento no desenvolvimento educacional, cientifico e tecnológico do país.

Justificativa

- (2) Existe uma carência de tecnologias voltadas para o desenvolvimento de soluções para a robótica Assistiva, além de aplicações de resgate e segurança;

- Portanto, devido a carência tecnologia de uma produção brasileira de robôs industriais e robôs de serviços, se faz necessário o estudo de métodos e técnicas que contribuam de alguma forma para suprir tal carência;

- No caso deste trabalho, temos como objeto de estudo a análise de expressões faciais de emoções básicas através do uso de uma adaptação de arquitetura de CNN - o LBCNN - que aplicada a plataformas robóticas autônomas possa trazer valor acadêmico e cientifico para a comunidade acadêmica brasileira além de atender demandas da sociedade em vários aspectos, tais como: auxilio doméstico e engenharia de defesa.

Objetivos

- Implementação de um sistema, capaz de realizar a tarefa de reconhecimento e classificação de expressões faciais básicas utilizando como extrator de características uma adaptação do LBCNN;

- Validação e análise entre os resultados obtidos (precisão de classificação, consumo computacional, etc.) em comparação a uma CNN tradicional.

Fundamentação Teórica

- Conceitos teóricos que dão embasamento para o desenvolvimento deste trabalho:
  - Redes Neurais Convolucionais;
  - Padrão Binário Local;
  - Reformulação do LBP através de filtros convolucionais.

Redes Neurais Convolucionais

- Uma CNN pode ser caracterizada como sendo uma NN de múltiplas camadas onde primariamente faz-se a suposição de que os dados de entrada sejam imagens;

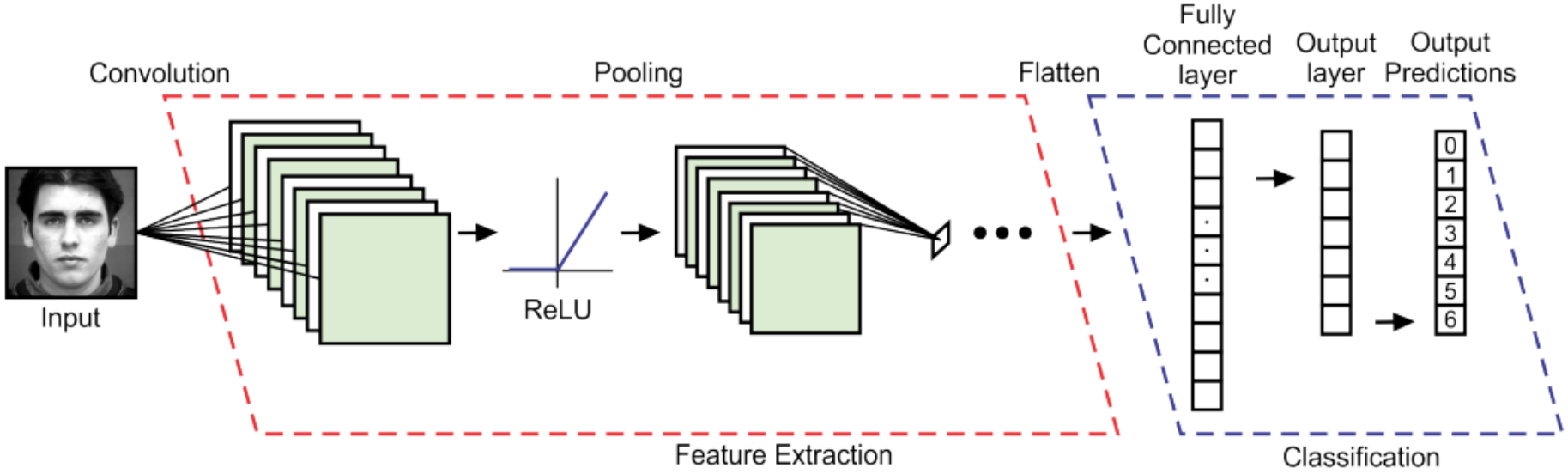
- Uma arquitetura básica de CNN pode ser dividida nas camadas listadas abaixo:

Redes Neurais Convolucionais

- Camada de Entrada;
- Camada de Convolução;
- Camada ReLU;
- Camada de Subamostragem;
- Camada de Flatten;
- Camada totalmente conectada.

Redes Neurais Convolucionais

Figura 4: Arquitetura básica do modelo CNN.



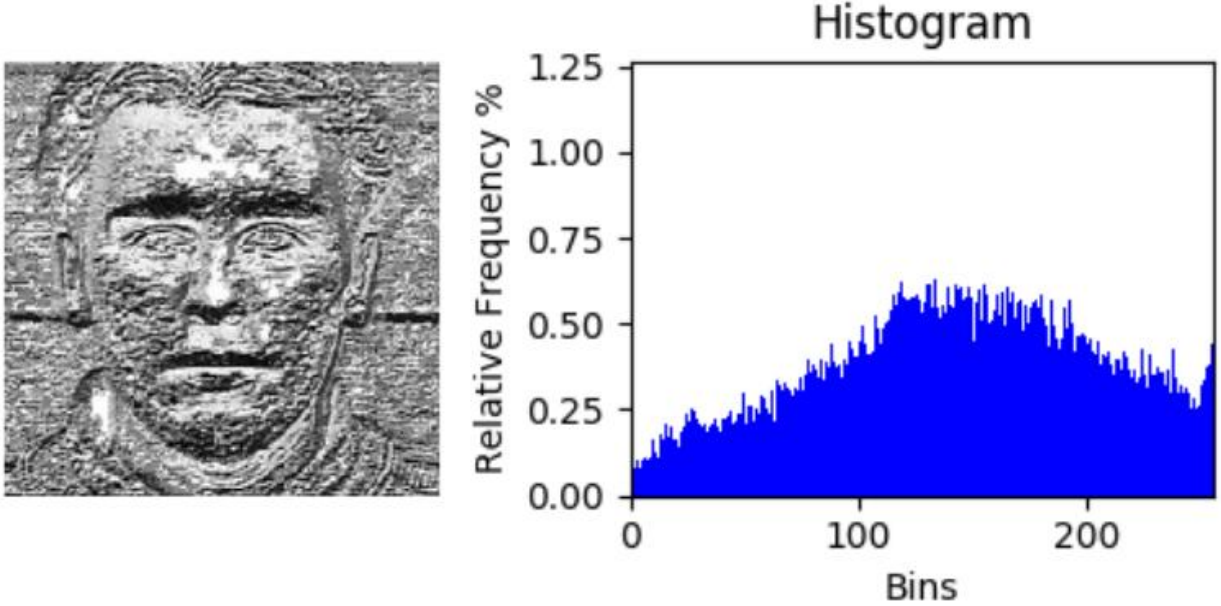
Fonte: Adaptado de [6].

Padrão Binário Local

- LBP é um tipo de descritor de textura simples e eficiente, usado para reconhecimento e classificação de padrões de imagens com texturas;
- Após diversas etapas, é então gerada o LBP resultante;
- Onde, posteriormente é calculado o histograma de intensidade dos pixels da imagem LBP resultante então gerada;
- Onde, o histograma 256-dimensional pode ser processado com algoritmos de ML (como por exemplo SVM) para realizar tarefas de análise de texturas e classificação de imagens.

Padrão Binário Local

Figura 7: Histograma de intensidade dos pixels.



Fonte: Elaborado pela autora.

Reformulação do LBP através de filtros convolucionais

- Para reformular o LBP através de filtros convolucionais e atingir o mesmo objetivo, é aplicado uma convolução de toda a imagem com 8 filtros convolucionais de tamanho 3x3, seguida de um operador de binarização não-linear, neste caso, a função Heaviside:

$$H(x) = \frac{1 + sgn(x)}{2}$$

Reformulação do LBP através de filtros convolucionais

- Onde ***sgn(x)*** se refere a função sinal, portanto, podemos obter a seguinte equação:

$$H(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x = 0 \\ 1, & x > 0 \end{cases}$$

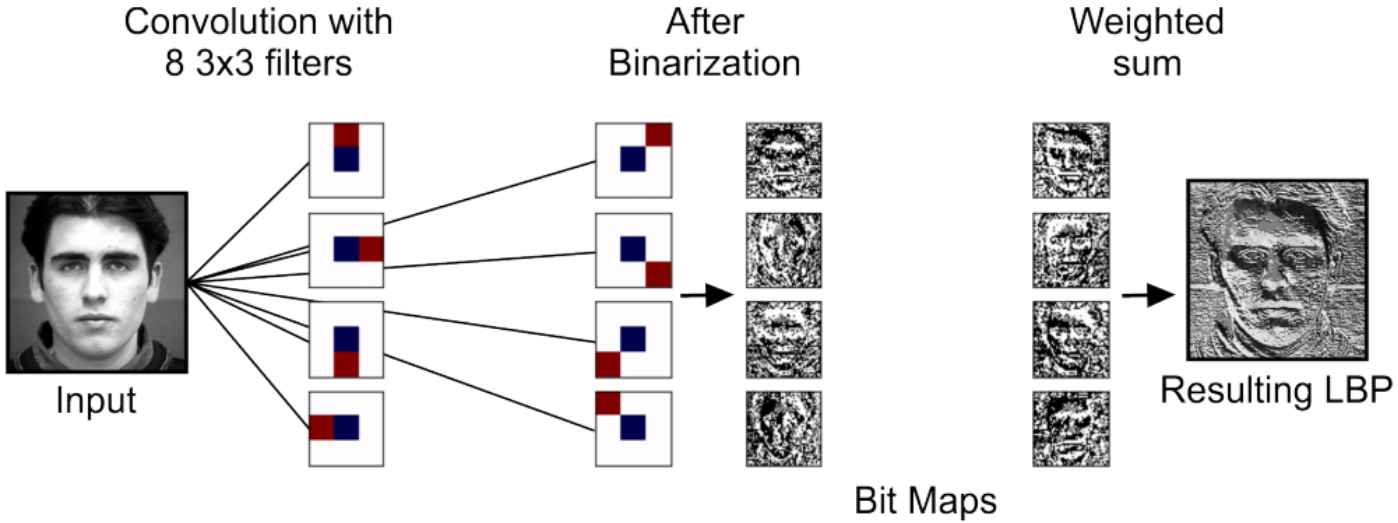
Reformulação do LBP através de filtros convolucionais

- Em seguida é realizada uma soma ponderada dos 8 bit maps usando um vetor de pesos pré-definidos, v = [2^7, 2^6, 2^5, 2^4, 2^3, 2^2, 2^1 , 2^0], onde são então extraídos os recursos LBP;

- A Figura 8 a seguir ilustra todo este processo descrito acima:

Reformulação do LBP através de filtros convolucionais

Figura 8: Reformulação do LBP através de filtros convolucionais.



Fonte: Adaptado de [3].

Reformulação do LBP através de filtros convolucionais

- Como mostrado na Figura 8, os filtros de intensidade vermelha indicam que o filtro terá uma resposta positiva aos pixels pretos nas imagens de entrada (1), enquanto os filtros de intensidade azul indicam que o filtro terá uma resposta negativa aos pixels pretos nas imagens de entrada (-1).

Implementação

- A adaptação do LBCNN proposta neste trabalho, bem como o pré-processamento das bases de imagens utilizadas, foi implementada em Python utilizando o framework TensorFlow em uma arquitetura baseada em GPU;

- Todos os experimentos realizados foram processados por meio de uma instância de máquina virtual de alto desempenho na infraestrutura do Google - Google Compute Engine (GCE) – um componente Infraestrutura como Serviço (em inglês, *Infrastructure as a Service* - IaaS).

Implementação

- O GCE é um recurso disponível na plataforma Google Cloud Platform (GCP);

- A instância de máquina virtual on demand é composta por uma GPU NVIDIA® Tesla® K80 com memória da GPU de 12GB GDDR5 e 8 vCPUs disponíveis, além de 16 GB de memória RAM e 50GB de HD e é acessada por meio da interface de linha de comandos.

Implementação

- Todos os experimentos foram realizados através da versão gratuita estendida do Google Cloud Platform por 12 meses;

- Onde ao criar uma conta na plataforma, é inserido um crédito de $ 300,00 dólares que poderá ser utilizado em todos os serviços oferecidos pela plataforma.

Implementação

GPU	-
Memória	12 GB de RAM
Armazenamento	500 GB de HD
Tempo gasto*	16 horas**

GPU	NVIDIA® GeForce® GTX 460 com memória de GPU de 1GB GDDR5.
Memória	8 GB de RAM
Armazenamento	500 GB de HD
Tempo gasto*	6 horas**

GPU	NVIDIA® Tesla® K80 com memória da GPU de 12GB GDDR5 e 8 vCPUs disponíveis.
Memória	16 GB de RAM
Armazenamento	50 GB de HD
Tempo gasto*	46 minutos**

*Etapa de treinamento. **Em média.

Base de imagens

- Foram utilizadas três bases de imagens públicas de expressão facial: JAFFE, Extended CK+, e FER-2013;

- As três bases de imagens possuem as seis expressões faciais básicas apresentadas, e neutro, logo, formando sete emoções a serem classificadas pelo nosso modelo de LBCNN.

Base de imagens

Figura 9: Amostras presentes nas três bases de imagens.

Fonte: Adaptado de [7, 8, 9, 10].

Pré-processamento

- Devido à natureza escassa das bases de imagens utilizadas neste trabalho, aplicamos o processo de Data Augmentation;

- Processo este que é aplicado ao conjunto de treinamento com a finalidade de gerar sub-amostras diferentes para cada imagem.

Figura 10: Resultado do processo de Data Augmentation.

Fonte: Elaborado pela autora.

Formato .pkl

- Pickle ou pickling (processo de serialização) é o termo empregado para o método de conversão aplicado a qualquer tipo de estrutura de objetos em Python para byte streams (0s e 1s);

- Na linguagem Python, existem diferentes módulos para a serialização de dados, Pickle é um deles;

- As vantagens em executar a serialização de dados está na possibilidade de salvar dados complexos, além da geração de algum nível de segurança de dados (mesmo que pequeno), uma vez que os dados serializados tornam-se quase ilegíveis.

Codificação One-Hot

- A codificação One-Hot é uma codificação de estados que consiste em gerar n vetores binários para n valores numéricos que representam cada um dos rótulos categóricos;

- Codificação necessária, pois:
  - (1) As CNNs não operam diretamente com rótulos categóricos;

  - (2) E a utilização de valores numéricos para representar rótulos categóricos em algoritmos de ML não é considerado como sendo uma boa prática, pois é possível que estes valores gerados possam prejudicar e influenciar a efetividade do algoritmo através do processo de aprendizado.

Codificação One-Hot

Tabela 1: Valores Numéricos x Codificação One-Hot.

Valores Numéricos	Codificação One-Hot
0	1000000
1	0100000
2	0010000
3	0001000
4	0000100
5	0000010
6	0000001

Rede Neural Convolucional Binária Local

- A fim de reduzir a complexidade computacional em CNN convencionais, o LBCNN se baseia nos princípios de LBP, e tráz como proposta, Local Binary Convolution (LBC), tornando- se uma poderosa alternativa para a camada convolucional em CNN convencionais;

- Além de reduzir cálculos, e significativamente a quantidade de parâmetros aprendíveis durante a etapa de treinamento devido a sua natureza binária e esparsa, a camada LBC reduz a complexidade do modelo, consequentemente, acarretando em economias computacionais e requisitos de memória, tornando-se um modelo aplicável em ambientes reais que possuem recursos escassos e limitados.

Rede Neural Convolucional Binária Local

- A camada LBC é composta por, (1) um conjunto de filtros convolucionais de caráter binário, pré-definidos e fixos, ou seja, durante a etapa de treinamento, não são atualizados; (2) seguida por uma função de ativação não linear (ReLU); (3) que é seguida, por um conjunto de pesos lineares 1x1 aprendíveis;

- Nosso modelo constitui-se em cinco camadas LBC, podendo ser ou não, seguida por max-pooling.

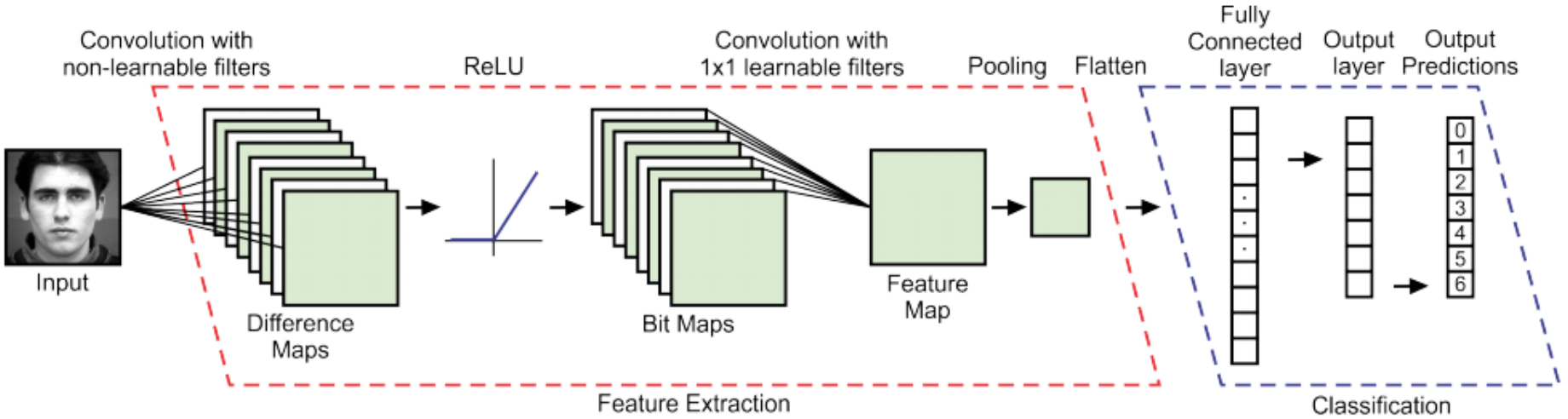
Rede Neural Convolucional Binária Local

- Utilizamos a distribuição de Bernoulli como uma generalização dos pesos em um LBP tradicional para gerar aleatoriamente nosso conjunto de filtros convolucionais de caráter binário, pré-definidos e fixos;

- Para isto, definimos para ser utilizado em todos os nossos experimentos, um nível de sparsidade igual a 0.5, em relação aos pesos que podem tolerar valores distintos de zero, em seguida atribuímos aleatoriamente 1 ou -1 a esses pesos.

Rede Neural Convolucional Binária Local

Figura 11: Arquitetura básica do modelo Local Binary Convolutional Neural Network.



Fonte: Adaptado de [3].

Experimentos e Resultados

- A mesma arquitetura utilizada para implementar nosso modelo de CNN convencional serviu como base para implementarmos também nosso modelo de LBCNN;

- Portanto, para fazer uma comparação junta entre as arquiteturas, todos os experimentos foram executados por 100 épocas, com um total de 20.000 iterações;

- Utilizamos uma taxa de aprendizado de ***1e-3***, e Adam Optimizer.

Experimentos e Resultados

- Desta forma, o número de filtros convolucionais, o número de camadas convolucionais, o número de unidades escondidas na camada totalmente conectada, foram os mesmos em ambos os modelos de redes;

- Já em relação aos parâmetros do LBCNN, utilizamos filtros de tamanho 3x3 para gerar os filtros não aprendíveis como no LBP tradicional, por apresentar melhores resultados nos experimentos realizados, além de serem gerados aleatoriamente pela distribuição de Bernoulli.

Experimentos e Resultados

- Os resultados obtidos através dos experimentos realizados podem ser vistos nas Tabelas 2 e 3;

- Onde a Tabela 2 mostra a acurácia obtida em cada base de imagem, enquanto a Tabela 3 mostra o tempo gasto na etapa de treinamento em cada base de imagem;

- As Figuras 12 e 13, respectivamente, mostram as matrizes de confusão das acurácias obtidas na etapa de teste do modelo CNN convencional e LBCNN implementadas neste trabalho.

Experimentos e Resultados

Tabela 2: Taxa de acurácia (porcentagem) na etapa de teste.

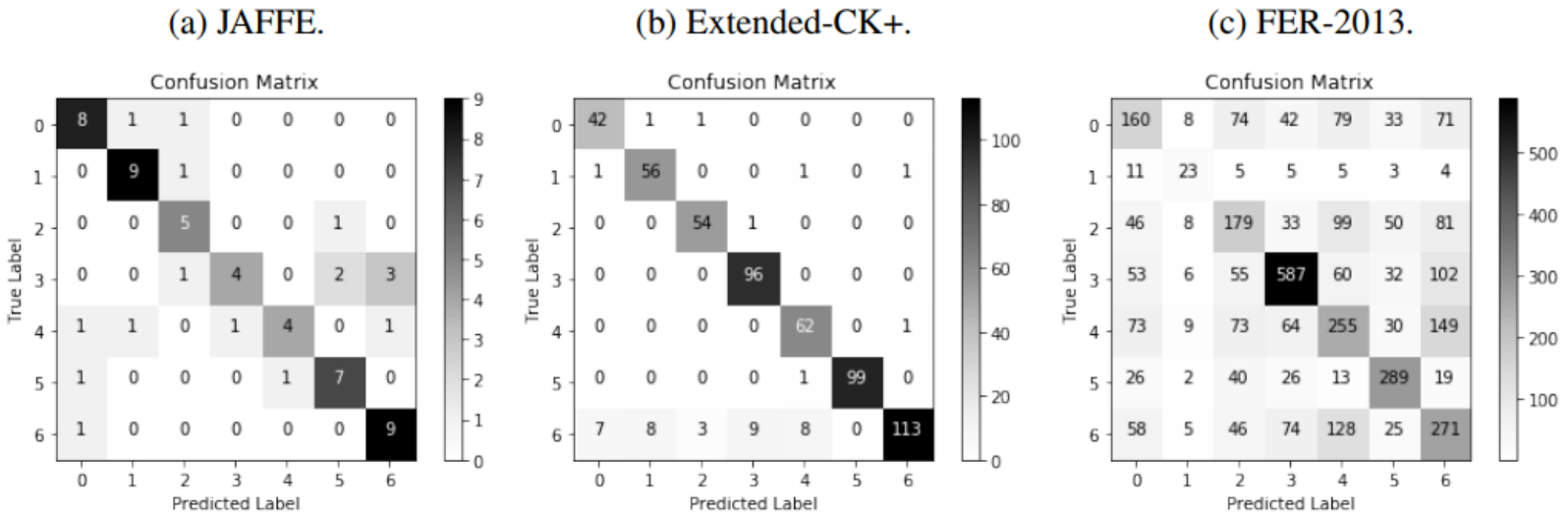
	CNN	LBCNN
JAFFE	73.0	77.8
Extended CK+	**92.4**	**82.1**
FER-2013	49.2	41.9

Tabela 3: Tempo gasto (horas) na etapa de treinamento.

	CNN	LBCNN
JAFFE	**0:46:15**	0:16:52
Extended CK+	0:46:23	0:16:55
FER-2013	0:46:31	**0:04:52**

Experimentos e Resultados

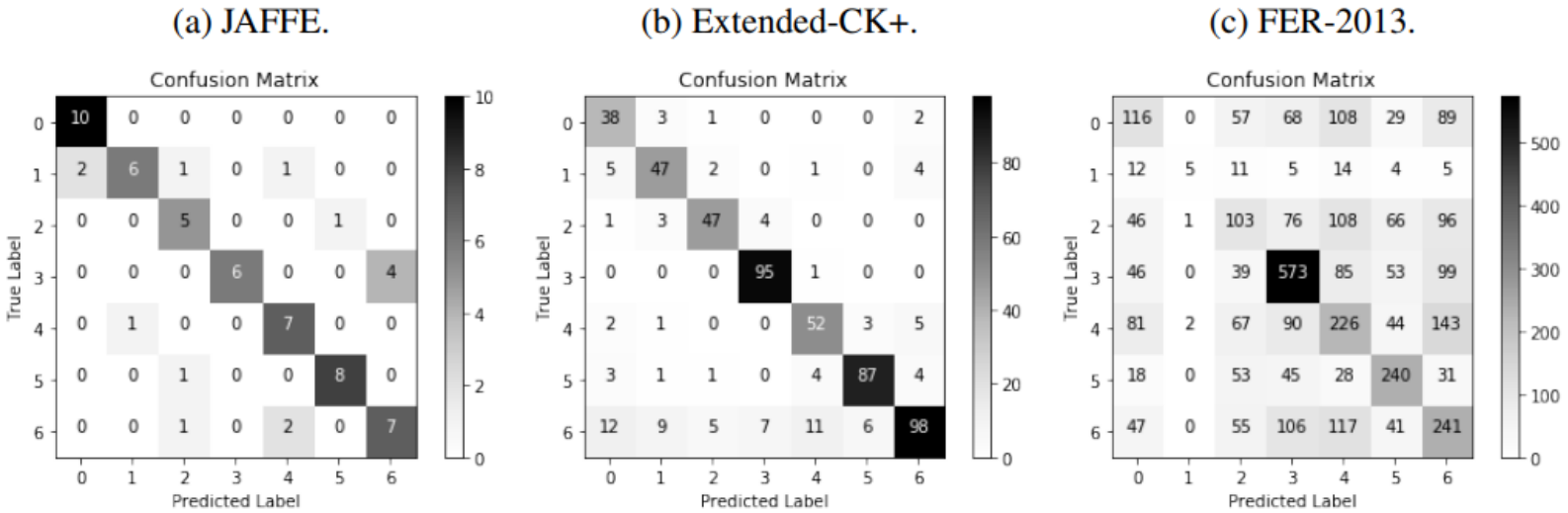
Figura 12: Matrix de Confusão da CNN convencional.



Fonte: Elaborado pela autora.

Experimentos e Resultados

Figura 13: Matrix de Confusão do LBCNN.



Fonte: Elaborado pela autora.

Experimentos e Resultados

- Além disto, foi estimado no GCE o seu custo efetivo ao utilizar a instância de máquina virtual de alto desempenho, onde o custo efetivo (em dólar) gerado ao executar ambos os modelos apresentados pode ser visto na Tabela 3 a seguir:

Tabela 3: Custo efetivo ao executar o modelo LBCNN em comparação ao modelo de CNN convencional.

	Custo	Tempo
CNN	0.639	46
LBCNN	**0.222**	**16**

Experimentos e Resultados

- Desta forma, é notável a eficiência e baixa complexidade computacional do LBCNN implementado neste trabalho;

- Sendo possível obter uma taxa de classificação satisfatória nos testes realizados em cada base de imagem com um custo baixo, gerando uma economia de $ 0.416 ao executar o modelo LBCNN em comparação ao modelo de CNN convencional.

Experimentos e Resultados

- Ressaltamos que o objetivo deste trabalho não é obter uma maior precisão na etapa de teste em relação à CNN convencional, mas sim revelar sua eficiência, economia e baixa complexidade computacional em relação à CNN convencional para a tarefa de FER que resulta em um tempo gasto na etapa de treinamento relativamente baixa em relação a CNN convencional, enquanto a taxa de acurácia obtida é, em vezes relativamente melhor ou insignificantemente baixa (se comparado ao tempo gasto na etapa de treinamento).

Considerações Finais

- Neste trabalho, propusemos o uso de LBCNN para a tarefa de FER, portanto, classificar as sete expressões faciais (Raiva, Nojo, Medo, Feliz, Triste e Surpresa) acrescentadas pela expressão neutra, compondo sete emoções básicas;

- O LBCNN foi implementado em Python usando o framework TensorFlow;

- Experimentos foram realizados para comparar sua eficiência com um modelo CNN convencional;

- Nossa abordagem mostrou-se eficiente em relação à sua precisão, custo efetivo e ao tempo gasto na etapa de treinamento, na qual é possível realizar a extração de características mais rapidamente.

Trabalhos Futuros

- Como continuação deste trabalho, propomos incorporar o modelo LBCNN apresentado neste trabalho em uma plataforma robótica autônoma;

- A plataforma robótica consiste de um robô Pioneer 3DX equipado com uma câmera RGBD, um laser de sensor Sick Lms200 e um computador usando o sistema operacional do robô (ROS), como descrito em [11, 12].

Trabalhos Futuros

- Ainda como continuação, pretendemos aplicar outras técnicas não abordadas neste trabalho, a fim de tornar o modelo de LBCNN proposto mais eficiente, tais técnicas como:

  - Validação cruzada (em inglês, Cross-validation) na etapa de treinamento a fim de ajudar a prevenir Overfitting;

  - Precision-Recall para Multiclass;

Trabalhos Futuros

- Regularização Dropout para, durante a etapa de treinamento regularizar as camadas totalmente conectadas da rede a fim de ajudar a prevenir Overfitting;

- Utilização do conjunto de ferramentas de visualização TensorBoard [89] para a otimização e depuração do framework TensorFlow.

Referências

- [1] Kanade, Takeo, Yingli Tian, and Jeffrey F. Cohn. "Comprehensive database for facial expression analysis." *fg*. IEEE, 2000.

- [2] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010.

- [3] Juefei-Xu, Felix, Vishnu Naresh Boddeti, and Marios Savvides. "Local binary convolutional neural networks." *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. Vol. 1. IEEE, 2017.

Referências

- [4] International Federation of Robotics. Executive SummaryWorld Robotics 2017 Industrial Robots. 2017. URL: https://ifr.org/downloads/press/Executive_Summary_WR_2017_Industrial_Robots.pdf (acesso em 11/06/2018).

- [5] International Federation of Robotics. Executive Summary World Robotics 2017 Service Robots. 2017. URL: https://ifr.org/downloads/press/Executive_Summary_WR_Service_Robots_2017_1.pdf (acesso em 11/06/2018).

- [6] Maurice Peemen, Bart Mesman e Henk Corporaal. “Efficiency optimization of trainable feature extractors for a consumer platform”. Em: International Conference on Advanced Concepts for Intelligent Vision Systems. Springer. 2011, pp. 293–304

Referências

- [7] Takeo Kanade, Jeffrey F Cohn e Yingli Tian. “Comprehensive database for facial expression analysis”. Em: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE. 2000, pp. 46–53.

- [8] Patrick Lucey et al. “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression”. Em: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE. 2010, pp. 94–101.

- [9] Michael Lyons et al. “Coding facial expressions with gabor wavelets”. Em: Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE. 1998, pp. 200–205. REFERÊNCIAS 65

Referências

- [10] Ian J Goodfellow et al. “Challenges in representation learning: A report on three machine learning contests”. Em: International Conference on Neural Information Processing. Springer. 2013, pp. 117–124.

- [11] Angonese, Alberto Torres, and Paulo Fernando Ferreira Rosa. "Integration of people detection and simultaneous localization and mapping systems for an autonomous robotic platform." *Robotics Symposium and IV Brazilian Robotics Symposium (LARS/SBR), 2016 XIII Latin American*. IEEE, 2016.

- [12] Angonese, Alberto Torres, and Paulo Fernando Ferreira Rosa. "Multiple people detection and identification system integrated with a dynamic simultaneous localization and mapping system for an autonomous mobile robotic platform." *Military Technologies (ICMT), 2017 International Conference on*. IEEE, 2017.