Deep Learning no Desafio Kaggle : Facial Expression Recognition Challenge

Jairo Lucas – Pedro Hoppe

Universidade Federal do Espirito Santo - Laboratório de Computação de Alto Desempenho

Na ultima década a análise de expressões faciais por um sistema autônomo tornou-se uma área de pesquisa bastante ativa em função do seu grande potencial de aplicações, visto que as expressões faciais refletem não somente as emoções, mas também outras atividades mentais e sinais fisiológicos.

A tarefa de reconhecimento de emoções por um sistema autônomo é uma tarefa extremamente complexa. Para se ter ideia deste nível de complexidade, podemos comparar esta tarefa com outra igualmente complexa, o reconhecimento facial. Enquanto alguns trabalhos de reconhecimento facial, como os de Schroff [7] e Sun Yi [8] já reportam uma acurácia acima de 99% para bases de imagens em ambiente não controlado, o ganhador do desafio de reconhecimento de emoções promovido na 16ª. International Conference on Multimodal Interaction [9] em 2014 obteve uma acurácia em torno de 50% para a base SFEW (Static Facial Expressions Wild) [12], base composta com imagens em ambiente não controlado. Na conferência do ano seguinte [10], Zhiding [11] conseguiu uma acurácia em torno de 61%.

No desafio "Challenges in Representation Learning: Facial Expression Recognition Challenge" [15] promovido pela Kaggle, Tang[17] se consagrou vencedor com uma acurácia de 71,16% .

Neste trabalho, nos propomos a criar uma Deep Learning CNN, seguindo os mesmo protocolos propostos pelo desafio Kaggle, e avaliar o desempenho da mesma em relação as soluções submetidas no concurso.

Para isso, utilizamos as bases de treino, teste e validação fornecidas pelo desafio, sem nenhum tipo de alteração nas mesmas, como a inclusão de novas imagens ou manipulação de imagens existentes. Os resultados obtidos foram bastante animadores, conseguimos uma taxa de torno de 55.3% de acurácia, oque nos colocaria em 19º lugar no referido desafio.

Termos –Kaggle, Redes Neurais Convolucionaiss, CNN, Deep Learning, Reconhecimento de expressões faciais, Reconhecimento de faces, Caffe.

1.INTRODUÇÃO

No final da década de 70 um grupo de pesquisadores descobriu que para um conjunto de emoções básicas existem expressões não-verbais distintas, universais, e provavelmente inatas. Desde então várias áreas de pesquisas se dedicaram a testar esta tese da universalidade, sendo hoje aceita a ideia que uma expressão de raiva, por exemplo, teria as mesmas características faciais no Brasil, na Europa ou na Ásia.

Um sistema computacional que pudesse avaliar corretamente, e em tempo real, as expressões faciais e o estado emocional do seu usuário, elevaria a interação entre homem e máquina para outro nível, diminuindo ou eliminado qualquer barreira de linguagem e permitindo uma interação ativa do sistema computacional com o seu usuário.

Essa possibilidade fez com que o interesse pelo tema aumentasse muito na última década, principalmente nas áreas de inteligência artificial e visão computacional.

1.2 – Técnicas de *Deep Learning*

As técnicas baseadas em *Deep Learning* – ganharam destaque entre os pesquisadores de várias áreas da inteligência artificial, principalmente aquelas voltadas para visão computacional. A grande maioria destas técnicas consiste em um conjunto de múltiplas tarefas de aprendizado de máquina que lida com diferentes tipos de abstrações. A técnica utilizada neste trabalho foi a de Redes Neurais Convolucionais – CNN.

1.2.1 – Redes CNN – Redes Neurais Convolucionais

Com a popularização dos celulares com câmaras digitais e a explosão de redes sociais como o Facebook e Instagram, o volume de informações em formato de imagens, fotos e vídeos disponíveis na internet vem aumentando de forma exponencial. O crescimento deste volume de informação demanda a criação de novas técnicas de buscas que não sejam baseadas em texto, e sim capazes de inferir informações diretamente destas mídias. As redes convolucionais foram projetadas para atuar neste tipo de problema, que envolve a detecção e reconhecimento de objetos, pessoas ou animais em uma determinada cena.

As CNN's – Redes Neurais Convolucionais - foram projetadas inspiradas na arquitetura biológica do celebro. Em 1968 Hubel e Wiesel realizaram experimentos com gatos e macacos e mostraram que o córtex visual é formado por um conjunto hierárquico de células sensíveis a pequenas sub-

regiões chamadas de campos receptivos, de forma que cada célula é "especialista" em monitorar (e ser ativada) por uma pequena região. Hubel classificava estas células em categorias - Simples, complexa e supercomplexa – de acordo com o padrão de estimulo que a ativam. Células simples são ativadas quando são apresentados padrões simples para o animal, como linhas. As células complexas e supercomplexas são ativadas quando padrões mais elaborados são apresentados ao animal.

A partir deste estudo surge a hipótese que uma boa representação interna para uma rede neural para reconhecimento de imagens seria uma estrutura hierárquica, onde os pixels formam arestas, as arestas formam padrões, os padrões combinados formam as partes, as partes combinadas formam os objetos e os objetos formam a cena. [4]

Esta estrutura considera que o mecanismo de reconhecimento necessita de vários estágios de treinamento empilhados uns sobre os outros, um para cada nível de hierarquia. As redes CNN's seguem este conceito, representando arquiteturas multi-estágios capazes de serem treinadas.

1.3 – Caffe - *Convolutional Architecture for Fast Feature Embedding*

O framework utilizado neste trabalho é o Caffe - *Convolutional Architecture for Fast Feature Embedding,* que foi projetado e desenvolvido pelos pesquisadores do *Berkeley Vision and Learning Center* (BVLC) da Universidade da California.

O Caffe é um framework de código aberto que oferece uma série de modelos e exemplos de redes pré-treinadas com Deep Learning. Possui uma comunidade bastante ativa [2] onde podem ser disponibilizados novos modelos e conhecimentos relevantes. Os modelos disponibilizados podem ainda ser adaptados ou parametrizados para uso em diversas aplicações.

Neste trabalho utilizamos o modelo *Vgg_face ,* modelo treinada originalmente para a tarefa de reconhecimento facial.

2. Trabalhos Relacionados

Este trabalho tem como foco a identificação de expressões faciais em imagens estáticas. Os resultados dos trabalhos nesta tarefa dependem muito da base de imagens utilizada. Trabalhos mais antigos usavam bases com imagens adquiridas em ambiente controlado, são imagens in-door, sem fundo complexo, em poses frontais e apenas uma face por foto. Para este tipo de base de imagens a acurácia relatada é próxima da perfeição como relatado no trabalho de Zhao[13] e no trabalho de Katsia e Pitas[14].

Trabalhos mais recentes utilizam bases de imagens mais próximas do "mundo real". São imagens out-door, adquiridas em situações reais, com fundo complexo e algumas vezes com mais de uma de uma face por imagem. Neste tipo de base de imagens a acurácia é bem menor, embora tenha melhorado significativamente nos últimos anos. Entre os trabalhos que utilizam este tipo de bases de imagens podemos citar Levi [1] que utilizou Redes CNN com mapeamento de padrões binários obtendo uma acurácia de 54% para a base SFEW, e Zhiding[11] que utiliza múltiplas redes Deep Learning e obtém um acurácia de 61% para esta mesma base.

Para a base de imagens FER-2013, utilizada neste trabalho, Tang[15], utilizou Deep Learning com SVM (Suport Vector Machines) conseguindo vencer o desafio "*Challenges in Representation Learning: Facial Expression Recognition Challenge"[16]* com uma acurácia 71.16% .

3.Metodologia,

Neste trabalho foram utilizadas as seguintes metodologias e ferramentas.

3.1 – Software

- Sistema operacional Linux Ubuntu
- Biblioteca Open MP
- Linguagem de programação C
- FrameWork Caffe

3.2– Hardware

GPU

Nvidia – Modelo "Tesla C2050"

Versão do Driver	:	7.5 / 7.5
Compatibilidade	:	2.0
Total memoria Global	:	2687 MBytes
14 SM com 32 Cores cada	:	448 CUDA Cores
Clock dos processadores	:	1147 MHz (1.15 GHz)
Memory Bus Width	:	384-bit
Total de memória	:	65536 bytes
Registradores por bloco	:	32768
Tamanho do Warp	:	32
Máximo de threads/bloco	:	1024
Máximo de threads/SP	:	1536

CPU

Intel Xeon

Número de cores	:	4 cores
Clock dos processadores	:	2.1 Ghz
Memória Ram	:	12 gb

3.3 – Base de Imagens

Este trabalho usou como base de imagens a FER-2013 [17], disponibilizada na página da competição. A base é composta por 35.887 imagens em tons de cinza, com tamanho de 48 x 48 pixels. As faces geralmente estão centralizadas e ocupam aproximadamente o mesmo espaço nas imagens. Existem imagens com poses laterais, parcialmente obstruídas com as mãos ou cabelos longos, com variação de iluminação e também imagens de bebês.

Todas as poses são classificadas em uma das seguintes categorias:

- 0 - Irritado
- 1 - Aborrecido / descontente
- 2 – Medo
- 3 – Feliz
- 4 – Triste
- 5 – Surpresa
- 6 – Neutro

A figura 1 mostra exemplos desta base de Imagens.

Fig.1 – Exemplos das expressões da vase Fer-2013

3.3.1 – Manipulação da entrada de dados

A base original de imagens fornecida no desafio é um arquivo do tipo CSV composto por três colunas delimitadas por virgula, onde a primeira coluna contém o código da expressão facial, a segunda coluna a sequencia de pixel que forma a imagem, e a terceira coluna informa onde a imagem deve ser utilizada (teste, treino ou validação).

Para adequar esta base para a entrada de nossa rede foi necessário criar uma rotina para ler a sequencia de pixels fornecida e criar uma imagem PNG desta sequencia. Para esta tarefa foi utilizada a linguem C em conjunto com a biblioteca Open CV.

3.3.2 – Conjuntos de Treino

Conforme o protocolo do desafio, o conjunto de treino é composto de um dataset com 28.709 imagens pré-definidas do conjunto principal. Neste conjunto estão representadas todas as categorias, conforme mostrado na figura 1.

3.3.3 – Conjuntos de Validação

O conjunto de validação, disponibilizado pela própria Kaggle, é composto por 3.589 imagens e utilizado para ajustes dos parâmetros da rede. As imagens utilizadas neste dataset não fazem parte do conjunto de treino e possuem representação das 7 categorias avaliadas.

3.3.2 – Conjuntos de Teste

O conjunto de Teste, disponibilizado pela própria Kaggle, é composto por 3.589 imagens e utilizado para aferir a acurácia final da rede utilizada. As imagens deste conjunto não fazem parte do conjunto de validação nem do conjunto de treino.

4. Experimentos e Resultados

Nos experimentos é apresentado ao classificador uma face e o mesmo deve classificar a expressão da face baseado nas expressões aprendidas na fase de treino da rede.

Foram utilizados dois experimentos, um para testar a acurácia no conjunto de validação e outro para testar a acurácia do conjunto de teste.

O resultado e a análise dos experimentos são descrito na sequencia.

4.1 – Resultado Para o Conjunto de Validação

Neste experimento foi utilizado o conjunto de validação, caracterizado no item 3.3.1, e efetuados diversos ajustes finos nos parâmetros da rede.

A figura 2 mostra o resultado obtido após a definição do melhor conjunto de parâmetros.

Conjunto de Treino : BaseTreinoII.csv
Conjunto de teste : BaseValidaçãoII.csv
Numero de iterações utilizada : 7.000
Tempo aproximado para o resultado : 0:45h

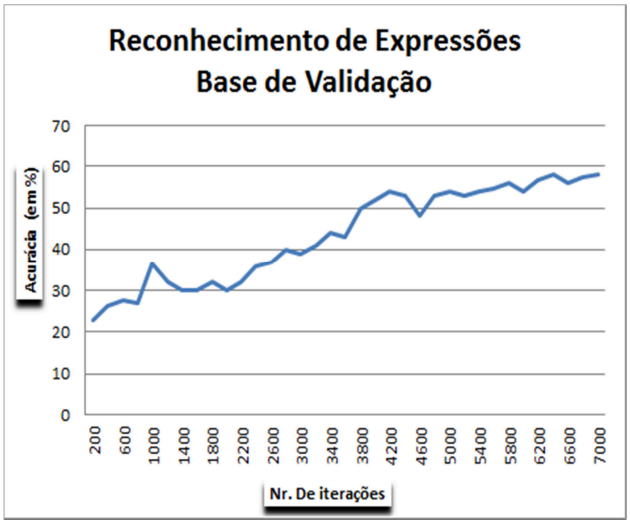

Fig.2. Resultados do experimento para o conjunto de Validação

O resultado mostra que após encontrar um conjunto bom de parâmetros a rede se mostrou razoavelmente estável, conseguindo um pico de acurácia de **58,3%.**

4.2 – Resultado Para o Conjunto de Teste

Neste experimento foi utilizado o conjunto de Teste caracterizado no item 3.3.2. Os parâmetros da rede que foram aferidos no experimento anterior foram aplicados neste experimento, sem nenhum tipo de alteração ou ajuste.

Na figura 3 é possível verificar os resultados obtidos para este conjunto.

Conjunto de Treino : BaseTreinoII.csv
Conjunto de teste : BaseTesteII.csv
Numero de iterações utilizada : 7.000
Tempo aproximado para o resultado: 0:45h

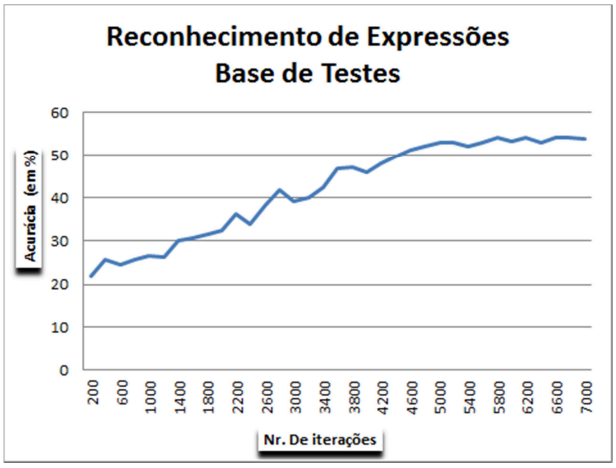

Fig.3. Resultados do experimento para o conjunto de Testes

Neste experimento a rede sofreu uma pequena queda na performance, oque já era esperado, pois a mesma sofreu um ajuste fino para a base de validação, sendo razoável perder um pouco de performance para uma base desconhecida.

A acurácia conseguida ficou em torno de **54%** após cerca de 7.000 iterações. Apesar de aparentemente modesta, esta acurácia nos colocaria entre os vinte primeiros colocados na competição em questão.

4 – Conclusões

Neste trabalho utilizamos uma rede CNN para a tarefa de reconhecimento de até 7 expressões faciais (Irritado, Aborrecido, Medo, Feliz, Triste, Surpresa, Neutro).

O experimento utilizou a base de imagens FER-2013[17], disponibilizada na competição "*Challenges in Representation Learning: Facial Expression Recognition Challenge"*[16], promovida pelo Kaggle. Esta base é formada por imagens de 48 x 48 pixel em tons de cinza, sendo uma única face por imagem, com a possibilidade de poses laterais, rosto parcialmente coberto pelas mãos ou cabelos. Além disso, a base conta com fotos de bebês e expressões sem muita definição, oque dificulta bastante à tarefa.

O estado da arte para esta base é do vencedor do concurso, Yichuan Tang[15], que conseguiu uma taxa de acurácia de **71.16%.** Na página do concurso não é descrita a técnica utilizada pelo vencedor, mas consultando alguns papers de autoria do mesmo, é possível concluir que foi utilizado Deep Learning com SVM (Suport Vector Machine). Para maiores detalhes, consultar[15].

Nossa rede conseguiu uma acurácia de aproximadamente **54%**, apesar de modesta, este número nos colocaria entre os vinte primeiros colocados do concurso. Nos testes com a base de validação, chegamos a obter um acurácia em torno de 58%, porém, conforme as regras da competição, a mesma não pode ser considerada para fins de comparação.

É provável que a acurácia obtida neste trabalho possa ser melhorada, pois tivemos que limitar muito nossa rede em função das limitações do hardware disponível. Em um primeiro momento utilizamos uma rede com 16 camadas, esta rede, com poucas iterações (cerca de 800), alcançava uma acurácia entre 30% a 35%, porém não conseguia prosseguir por falta de memória disponível. Diante disso fomos obrigados a reduzir gradualmente o número de camada da rede até que a mesma fosse suportada pelo hardware, oque com certeza, impactou no desempenho da mesma.

5 -BIBLIOGRAFIA

[1] Levi, Gil and Tal Hassner; ***Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns*** ; ACM International Conference on Multimodal Interaction (ICMI), Seattle, . 2015

[2] Jia, Yangqing and Shelhamer, Evan and Donahue, Jeff and Karayev, Sergey and Long, Jonathan and Girshick, Ross and Guadarrama, Sergio

and Darrell, Trevor; ***Caffe: Convolutional Architecture for Fast Feature Embedding***; arXiv preprint arXiv:1408.5093; 2014

[3] A. M. Martinez. **The AR face database,** CVC Technical Report, 1998,

[4] LECUN, Yann et al. ***Convolutional networks and applications in vision.*** In: ISCAS. 2010. p. 253-256.

[5] Kobayashi, H. and Hara, F. ***The Recognition of Basic Facial Expressions by Neural Network***. Proc. IJCNN 1991, IEEE Computer Society (1991), 460-466.

[6] Happy ,S L ; Routray, A; ***Automatic Facial Expression Recognition Using Features of Salient Facial Patches***; IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 6, N. 1, JANUARY-MARCH 2015

[7] Schroff , Florian; Kalenichenko, Dmitry and Philbin , James ; ***FaceNet: A Unified Embedding for Face Recognition and Clustering***. In Proc. CVPR, 2015.

[8] Sun ,Yi; Liang ,Ding; Wang, Xiaogang; Tang, Xiaoou ; ***DeepID3: Face Recognition with Very Deep Neural Networks***.; CoRR, abs/1502.00873, 2015

[9] Dhall, Abhinav; Goecke,Roland; Joshi, Jyoti; ***Emotion Recognition In The Wild Challenge 2014:Baseline, Data and Protocol***; (EmotiW), 2014.

[10] Dhall, Abhinav; Goecke,Roland; Murthy, O. V. Ramana; ***Emotion Recognition In The Wild Challenge 2015:Baseline, Data and Protocol***; (EmotiW), 2015.

[11] Zhiding Yu, Cha Zhang ; ***Image based Static Facial Expression Recognition with Multiple Deep Network Learning***; Wild Challenge (EmotiW), 2015.

[12] A. Dhall, R. Goecke, S. Lucey and T. Gedeon, ***Static Facial Expression in Tough Conditions - EFEW***: Data, Evaluation Protocol and Benchmark, IEEE ICCV BEFIT Workshop 2011

[13] Zhao, J., and G. Kearney. "***Classifying facial emotions by backpropagation neural networks with fuzzy inputs***" International Conference on Neural Information Processing. Vol. 1. 1996.

[14] Kotsia, Irene, and Pitas, Ioannis. "***Facial expression recognition in image sequences using geometric deformation features and support vector machines***." IEEE Transactions on Image Processing 16.1 (2007): 172-187.

[15] Tang, Yichuan. "***Deep Learning using Linear Support Vector Machines***". International Conference on Machine Learning; Challenges in Representation Learning Workshop. Atlanta, Georgia, 2013

[16] Kaggle. ***Challenges in Representation Learning: Facial Expression Recognition Challenge***. Disponível em: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge , acesso em 09/12/2016

[17] Kaggle. ***Challenges in Representation Learning: Facial Expression Recognition Challenge - DataBase***. Disponível em: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data , acesso em 09/12/2016