Classificação de Expressões Faciais com *Ensemble* de Redes Neurais Convolucionais e Votação Inteligente

Rodrigo C. Moraes, Carlos Maurício S. Figueiredo, Elloá B. Guedes

[1]Núcleo de Computação
Escola Superior de Tecnologia
Universidade do Estado do Amazonas
Av. Darcy Vargas, 1200 – Manaus – Amazonas

{rcm.eng, cfigueiredo, ebgcosta}@uea.edu.br

Abstract. *Facial Expression is a very important factor in the social interaction of human beings. And technologies that can automatically interpret and respond to stimuli of facial expressions already find a wide variety of applications, from anti-depressant drug testing to fatigue analysis of drivers and pilots. In this context, the following work presents a model for Automatic Classification of Facial Expression using as a training base the dataset Challenges in Representation Learning (FER2013), characterized by examples of spontaneous facial expressions in uncontrolled environments. The presented method is composed by a Convolutional Neural Networks Ensemble architecture, using a non-trivial voting system, based on a smart model, Xtreme Gradient Boosting - XGBoost. As performance criteria for validation of the proposed model, were used K-fold and F1 Score Micro techniques to guarantee robustness and reliability of the results, which are competitive with state-of-the-art works.*

Resumo. *A Expressão Facial é um fator de suma importância na interação social dos seres humanos. E tecnologias que podem interpretar e responder de forma automática a estímulos de expressões faciais já encontram uma grande variedade de aplicações, desde teste de fármacos anti-depressivos, até análise de fadiga de motoristas e pilotos. Neste contexto, o seguinte trabalho apresenta um modelo para Classificação Automática de Expressão Facial utilizando como base de treinamento o dataset Challenges in Representation Learning: Facial Expression Recognition Challenge(FER2013), caracterizado por exemplos de expressões faciais espontâneas em ambientes não controlados. O método apresentado é composto por uma arquitetura Ensemble de Redes Convolucionais Neurais, utilizando um sistema de votação não-trivial, baseado em um modelo inteligente, Xtreme Gradient Boosting - XGBoost. Como critérios de desempenho para validação do modelo proposto foram empregadas técnicas de K-fold e F1 Score Micro, para garantia de robustes e confiança dos resultados, que são competitivos com trabalhos estado-da-arte.*

Introdução

As expressões faciais são um componente não-verbal da comunicação humana, compreendendo sorrisos, franzir de sobrancelhas e outras ações, com vistas a transmitir emoções e sentimentos [Ekman and Friesen 1971]. Por meio do reconhecimento e interpretação de expressões faciais é possível obter um *feedback* mais verossímil e imediato da percepção dos indivíduos sobre um determinado contexto ou situação, razão pela qual é adotada, por exemplo, para avaliação do efeito de fármacos, detecção de fadiga de motoristas e pilotos,

etc. [Fasel and Luettin 2003]. Dada a importância desta percepção, é natural a demanda pelo desenvolvimento de soluções automáticas que capturem o significado das expressões faciais de indivíduos.

Nesta perspectiva, a classificação automática de expressões faciais consiste em localizar faces humanas em uma cena, extrair características faciais da região detectada, reconhecer padrões e então categorizar o resultado obtido em uma das sete expressões faciais canônicas: felicidade, tristeza, medo, nojo, surpresa, raiva e neutro [Pantic 2009]. Esta tarefa, entretanto, envolve a superação de diversos desafios, tais como as distintas formas que estas expressões podem ser manifestadas por indivíduos diferentes, a presença de outros elementos corporais, como as mãos, para composição da expressão, e até mesmo as características pessoais dos sujeitos, como presença ou ausência de barba, óculos, etc.

Com o advento dos métodos de *Machine Learning*, houve um progresso na classificação automática de expressões faciais. As técnicas de *Deep Learning*, em particular, têm colaborado para o avanço do estado da arte neste problema, este sucesso está relacionado a alta habilidade de representação das informações de entrada, utilizando várias camadas de neurônios artificiais compondo vários níveis de abstração[LeCun et al. 2015]. Porém, mais esforços precisam ser efetuados para uma melhor eficiência nesta tarefa.

Considerando o contexto apresentado, este trabalho se propõe a apresentar uma abordagem baseada no uso de Redes Neurais Convolucionais organizadas segundo um *ensemble* com votação mediada por um modelo de *Machine Learning* baseado em *Boosting*, para determinação das expressões correspondentes a partir de imagens de faces humanas. Os resultados obtidos com esta abordagem mostram-se animadores, pois obteve-se um desempenho de 71.74% nesta tarefa, equiparável com algumas das contribuições mais recentes da literatura neste problema.

Para apresentar os resultados obtidos, este trabalho está organizado como segue. A Seção X apresenta uma visão geral do estado da arte para a classificação automática de expressões faciais humanas com técnicas de *Deep Learning*. Em seguida, na Seção Y, são apresentados os materiais e métodos a serem considerados na elaboração da solução proposta. Os resultados obtidos e a discussão são então apresentados na Seção Z. Por fim, as considerações finais e perspectivas de trabalhos futuros são mostradas na Seção W. Completar

Trabalhos Relacionados

Relatar aqui os trabalhos análogos.

Materiais e Métodos

Dados Experimentais

A base de dados de expressões faciais utilizada para o desenvolvimento deste trabalho é denominada *Facial Expression Recognition Challenge* (FER2013). Esta base contém 35.887 imagens faciais em escala de cinza com dimensões de 48×48 pixels, rotuladas de maneira supervisionada segundo uma das sete expressões faciais universais, conforme amostras ilustradas na Figura 1.

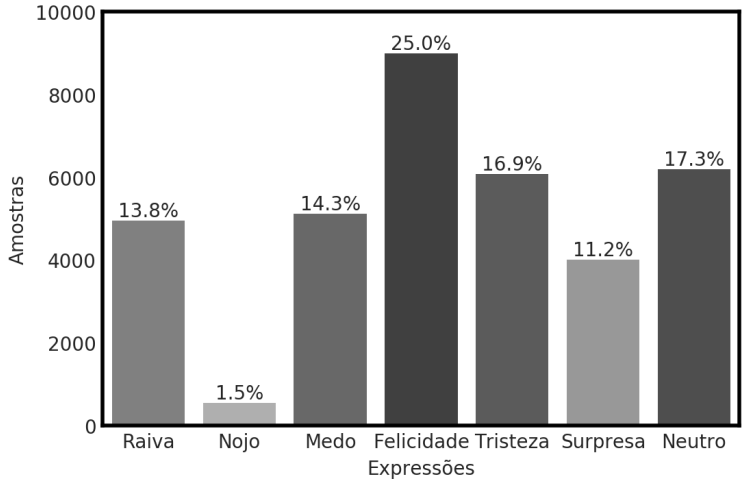
Conforme ilustra a Figura 1, é interessante notar algumas características particulares das imagens do FER2013 que ressaltam a relevância desta base de dados. Observa-se que, embora as faces estejam centralizadas nas imagens, elementos como cortes de cabelo, barba, óculos e até mesmo mãos encontram-se presentes, diminuindo a distância entre os exemplos contidos nesta base de dados e aqueles passíveis de ocorrência em um cenário realístico.

Figura 1: Amostras de imagens faciais da base de dados FER2013.

(a)Felicidade, (b)Tristeza, (c)Medo, (d)Nojo, (e)Surpresa, (f)Raiva, (g)Neutro.

Os exemplos disponíveis na FER2013 se distribuem de maneira heterogênea perante as classes consideradas, conforme ilustra o gráfico da Figura 2. O número de exemplos rotulado com a expressão "nojo", por exemplo, representam apenas 1.5% do total de exemplos disponíveis. Estas características evidenciam o desbalanceamento do conjunto de dados considerado no tocante à quantidade de amostras por classe.

Figura 2: Distribuição de imagens por tipo de expressão facial na base FER2013.

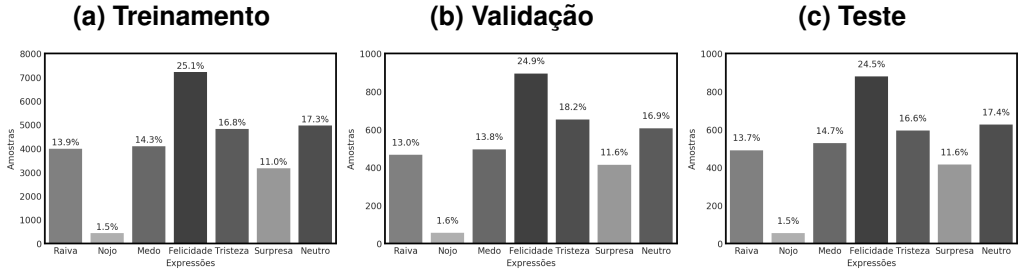


Descrição da Tarefa de Aprendizado de Máquina

A base de dados em questão será usada para realização de tarefa de classificação multirótulo segundo o paradigma de aprendizado supervisionado. Nesta tarefa, exemplos de expressões faciais e seus respectivos rótulos serão fornecidos previamente aos modelos de Aprendizado de Máquina para escolha e ajuste de parâmetros e realização do treinamento. Posteriormente, expressões faciais ainda não vistas serão apresentadas e o objetivo será avaliar o desempenho do modelo na classificação destes exemplos, isto é, aferir a respectiva capacidade de generalização.

Obedecendo a uma partição do FER2013 previamente considerada em competições de Visão Computacional [Kaggle 2013], esta base de dados será dividida em 3 partes, sendo: 75% dos exemplos para treinamento, $12,5\%$ dos exemplos para validação e $12,5\%$ dos exemplos remanescentes para testes, a serem utilizados seguindo uma abordagem de *holdout* de validação cruzada [Brink et al. 2017]. Apenas os exemplos da partição de testes serão utilizados para obtenção das métricas de desempenho e comparação dos modelos. Conforme ilustra a Figura 3, as partições preservam a distribuição de amostras por classe na base de dados original.

Figura 3: Distribuição das classes nas partições adotadas para o conjunto de dados.



Levando em conta os modelos de Aprendizagem de Máquina para a tarefa em questão, é essencial que um número razoável de exemplos esteja disponível para um ajuste apropriado dos parâmetros treináveis. Considerando esta necessidade prática, os exemplos da partição de treinamento passaram por um processo de pseudo-expansão do tipo *data augmentation*, em que novas imagens foram geradas a partir das previamente existentes considerando operações de rotação, espelhamento e

complementar aqui

, colaborando para a posterior regularização dos modelos [Chollet 2017]. Ao final desta etapa, o conjunto de treinamento passou a conter $1.3274E + 15$ exemplos, um aumento de $3.6989E + 10$ vezes em relação ao seu tamanho original, mas preservando a distribuição de exemplos nas classes.

completar aqui

Técnicas utilizadas no *augmentation*

- Rotação
- Descolocamento horizontal
- Descolamento vertical
- Escala
- Reflexão no eixo y

A métrica de desempenho adotada para comparação dos modelos na realização desta tarefa foi o Micro F-*Score*. Embora a acurácia seja uma métrica mais popular, que descreve o percentual de acertos do modelo em relação ao total de previsões efetuadas, não fornece detalhes acerca dos acertos por classe. Para contornar esta dificuldade, o Micro F-*Score* foi preferido, pois contempla a média harmônica entre precisão e revocação por classe ao passo que considera as diferentes frequências nas classes do problema [Kubat 2015]. Esta métrica é especialmente utilizada em problemas de classificação com classes desbalanceadas, ou seja, em situações análogas ao cenário considerado no escopo deste trabalho.

Dentre os modelos a serem avaliados, serão elencados como mais aptos para a tarefa de classificação proposta aqueles que maximizarem a métrica de desempenho Micro F-*Score* para os exemplos pertencentes à partição de testes.

Proposição de Modelos

Os modelos propostos e seus respectivos parâmetros e hiperparâmetros para a tarefa de classificação de expressões faciais são descritos detalhadamente ao longo desta seção.

Seguindo a abordagem predominantemente adotada pelo estado da arte no tocante ao aprendizado de características em dados de alta dimensionalidade para tarefas de Visão Computacional [Khan et al. 2018], as redes neurais convolucionais foram o modelo de

Aprendizado de Máquina adotado na tarefa elencada. Em particular, a arquitetura base foi a da rede neural convolucional canônica VGG-16 [Simonyan and Zisserman 2015], mas com algumas adaptações. Esta arquietura originalmente proposta destacou-se mediante a ideia de que uma rede neural precisa ter uma quantidade razoável de camadas convolucionais profundas para uma representação hierárquica adequada das informações visuais.

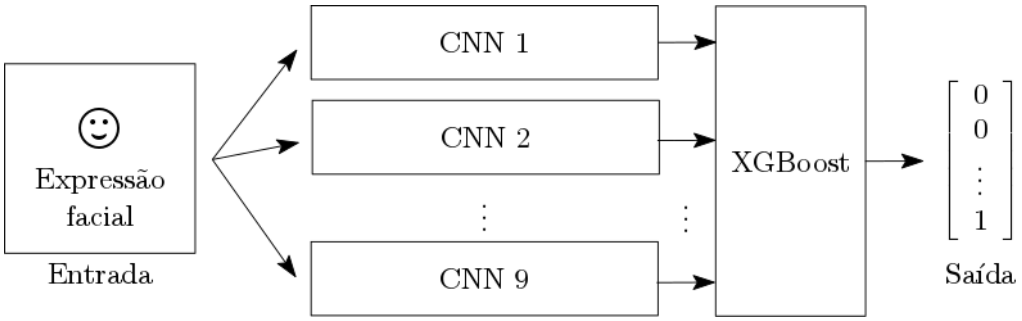
As adaptações da VGG-16 levaram em conta diferentes quantidades de repetições de certas operações, apresentadas de acordo com a seguinte ideia geral:

$$\begin{aligned} \text{Camada de Entrada} \Rightarrow [&(\text{Convolução} \rightarrow \text{Batch Normalization}) \cdot i \\ &\Rightarrow (\text{Pooling} \rightarrow \text{Dropout}) \cdot j] \cdot k \\ &\Rightarrow [\text{Fully Connected} \rightarrow \text{ReLu}] \cdot \ell \\ &\Rightarrow \text{Flatten} \Rightarrow \text{Camada de Saída}, \end{aligned} \tag{1}$$

em que i, j, k, ℓ são números inteiros que denotam a quantidade de repetições da operação associada perante multiplicação. Os valores destes inteiros foram: $i = 2$, $j = \{1, \ldots, 5\}$, $k = \{2, 3, 4\}$ e $\ell = \{1, 2, 3\}$. Considerando esta abordagem de adaptação, foram então propostas 9 CNNs diferentes a serem treinadas e testadas, conforme a abordagem de validação cruzada previamente descrita.

Além da avaliação individual do desempenho das CNNs propostas, considerou-se também a posterior combinação destas redes em um *ensemble*. Para valoração da classificação final, ao invés de considerar as abordagens típicas de votação unânime ou majoritária adotadas por *ensembles*, utilizou-se um modelo baseado em *Boosting*, o XGBoost [Chen and Guestrin 2016]. Este modelo recebe as saídas de todas as redes individuais e, após ter sido treinado com os exemplos da base de dados, decide dentre as classificações individuais qual a classificação final mais apropriada. Observa-se aqui uma modificação não-trivial em *ensembles*: a votação mediada por um modelo inteligente. A Figura X ilustra a ideia considerada.

Figura 4: *Ensemble* de CNNs com votação mediada por XGBoost.



Resultados e Discussões

O modelo proposto conseguiu superar os seres humanos nesta base de dados, na tarefa de classificar as expressões faciais, visto que sua acurácia foi de 71.74% enquanto a dos seres humanos é de $65 \pm 5\%$ [Goodfellow et al. 2013].

Os resultados de acurácia e *F1 Micro* para os modelos de CNN testados, bem como *Ensemble* obtiveram os mesmos valores, visto isso é apresentado somente o valor da medida *F1 Score Micro*. Na Tabela 1 observa-se os resultados para cada modelo de CNN, juntamente do resultado do *Ensemble*.

Tabela 1: F1 Micro das Arquiteturas utilizadas

Modelo	F1 Micro
1	0,6898857620507105
2	0,6767901922541097
3	0,6606297018668152
4	0,6798551128448036
5	0,6667595430482028
6	0,6781833379771524
7	0,6949010866536640
8	0,6285873502368348
9	0,6244079130677069
Ensemble	0,7174700473669546

É observado que o melhor classificador, modelo 7, individual de CNN, obteve 69.49% enquanto que o pior, modelo 9, obteve 62.44%. Ressaltando que cada classificador de CNN usado no *Ensemble* obteve resultados melhores do que os outros, em determinada expressão ou bons resultados em todas as expressões, mas não se sobressaiu em nenhuma expressão específica. No caso do modelo 9, obteve-se bons resultados em quase todas as expressões, mas nenhum resultado melhor na classificação de determinada expressão em relação aos outros modelos. Já no caso do modelo 7, obteve-se o melhor resultado de classificação para expressão de surpresa.

No modelo *Ensemble* teve-se um total de 32.738.871 parâmetros treináveis, que é resultado da soma dos parâmetros treináveis dos modelos CNN, bem como do XGBoost. Onde o modelo de XGBoost possui uma quantidade pequena de parâmetros treináveis, em comparação com os modelos de CNN, pois, enquanto a quantidade deste está na ordem de milhares, os outros estão na ordem de milhões. Este comportamento se mostra coerente, se for levado em consideração as tarefas de cada modelo, que possuem complexidades bastante distintas.

Tabela 2: Quantidade de parâmetros treináveis

Modelo	Qtd. Parâmetros Treináveis
1	15.250.887
2	1.778.247
3	2.134.471
4	1.778.247
5	1.189.511
6	1.780.103
7	6.653.599
8	1.186.231
9	923.575
XGBoost	64.000
Total	32.738.871

Os modelos de CNN mais profundos, consequentemente com maior número de parâmetros treináveis, foram os que obtiveram melhores desempenho individuais. Já os modelos mais rasos, apesar de obterem desempenho abaixo dos profundos, obtiveram os melhores resultados em classificar expressões específicas. Quantidade de parâmetros treináveis do *Ensemble* podem ser visualizados na Tabela 2.

Na Figura 5 é apresentado o resultado do modelo utilizando *Ensemble*. Onde o resultado final de desempenho foi de 71.74% de acordo com a métrica *F1 Micro*, ressaltando que seu valor de Acurácia possui o mesmo valor. E com este resultado o *Ensemble* supera, por pouco, o modelo campeão da competição.

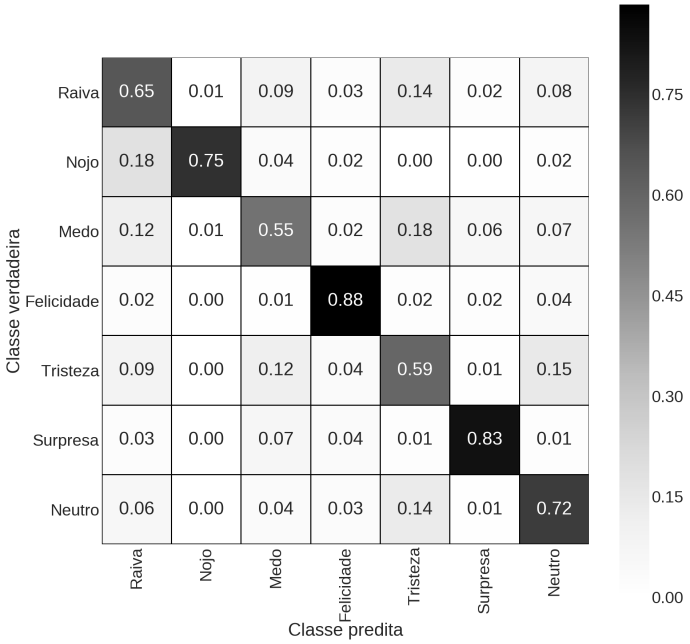


Figura 5: Matriz de Confusão do *Ensemble* (CNN + *XGBoost*)

Apesar do desbalanceamento da base de dados, a classificação da expressão de nojo obteve-se um dos melhores resultados 1, acompanhadas de surpresa e felicidade, com *F1 Score* de 78% 83% e 89% respectivamente. As classificações das outras expressões também apresentaram bons resultados, todas com *F1 Score* maior ou igual a 58%

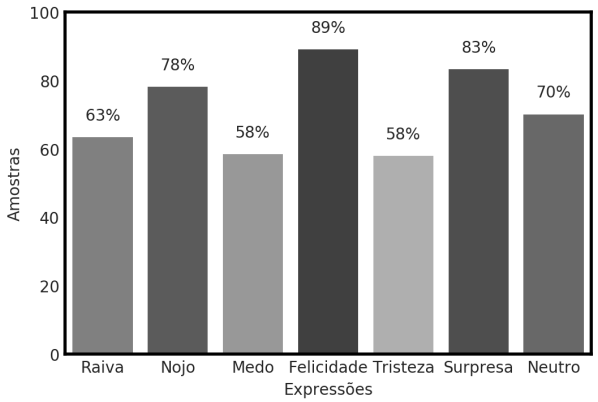


Figura 6: F1 Score por Expressão

Considerações Finais

Texto

Referências

[Brink et al. 2017] Brink, H., Richards, J. W., and Fetherolf, M. (2017). *Real-World Machine Learning*. Manning Publications, Estados Unidos.

[Chen and Guestrin 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 785–794, New York, NY, USA. ACM.

[Chollet 2017] Chollet, F. (2017). *Deep Learning with Python*. Manning Publications, Shelter Island, New York, 1 edition.

[Ekman and Friesen 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129.

[Fasel and Luettin 2003] Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275.

[Goodfellow et al. 2013] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer.

[Kaggle 2013] Kaggle (2013). Challenges in representation learning: Facial expression recognition challenge.

[Khan et al. 2018] Khan, S., Rahmani, H., Shah, S. A. A., and Bennamoun, M. (2018). *A Guide to Convolutional Neural Networks for Computer Vision*. Morgan and Claypool.

[Kubat 2015] Kubat, M. (2015). *An Introduction to Machine Learning*. Springer, Estados Unidos.

[LeCun et al. 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

[Pantic 2009] Pantic, M. (2009). *Facial Expression Analysis*, volume 6, pages 400–406.

[Simonyan and Zisserman 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, EUA.