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

A Classificação de Expressões Faciais é um processo executado por humanos e computadores que consiste em localizar faces em uma cena, extrair características faciais da região detectada, analisar alterações das características faciais como um sorriso ou um franzir de sobracelhas, e categorizar o resultado em uma expressão como felicidade ou raiva, por exemplo [Pantic 2009].

Tecnologias que podem interpretar e responder de forma automática a expressões faciais já encontram uma grande variedade de aplicações, dada sua importância social.

Exemplos disto, são sistemas de ensino que utilizam a expressão facial dos alunos como *feedback*, teste da efetividade de fármacos anti-depressivos e detecção de fadiga de motoristas e pilotos [Fasel and Luettin 2003].

Graças a introdução de métodos de *Machine Learning*, tem-se avançado no campo de Classificação Automática de Expressões Faciais. Mais especificamente, métodos de *Deep Learning* tem apresentado resultados bons nas tarefas que envolvem o uso de detecção de padrões e extração de características em imagens, nas mais variadas situações e contextos [Whitehill et al.].

Paul Ekman e Friesen postularão seis emoções primárias que possuem cada uma conteúdo próprio e associação a uma única expressão facial. Estas emoções se monstram invariantes ao longo das diversas culturas humanas e são identificadas como felicidade, tristeza, medo, nojo, surpresa e raiva [Ekman and Friesen 1971].

O presente trabalho apresentou resultados equiparáveis ao da literatura na tarefa de Classificação Automática de Expressões Faciais nas expressões primárias, utilizando modificação não-trivial, votação mediada por um modelo inteligente, em *Emsemble* de Redes Neurais Convolucionais.

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.

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

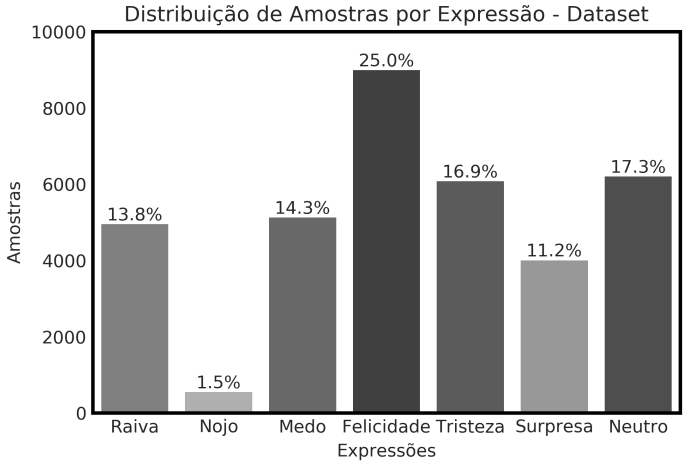
(a)Felicidade, (b)Tristeza, (c)Medo, (d)Nojo, (e)Surpresa, (f)Raiva, (g)Neutro.

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.

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: Histograma da distribuição de imagens por tipo de expressão facial na base FER2013.



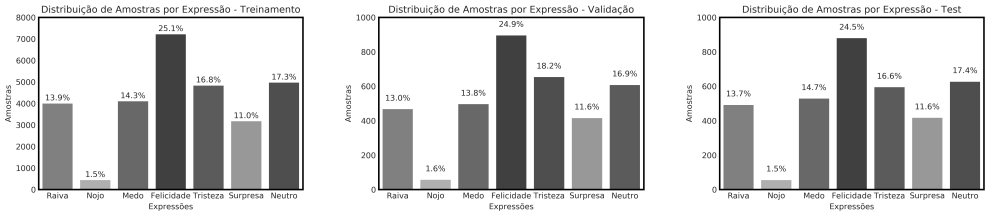
Descrição da Tarefa de Aprendizado de Máquina

A base de dados em questão será usada para realização de tarefa de classificação multi-rótulo 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.

(a) Treinamento **(b) Validação** **(c) Teste**



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

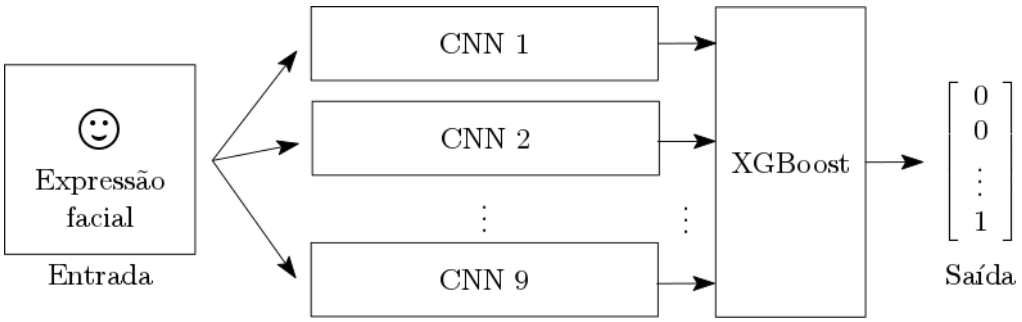
de certas operações, apresentadas de acordo com a seguinte ideia geral:

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

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

A métrica para cálculo de desempenho dos modelos durante a competição no *Kaggle* foi a acurácia, e o melhor resultado obtido na competição foi de 71.161%. Os resultados de acurácia e *F1 Micro* para os modelos de CNN testados, bem com o *Ensemble* obtiveram os mesmos valores, visto isso é apresentado somente o valor da medida *F*. Na Tabela 1 observa-se os resultados para cada modelo de CNN, juntamento do resultado do *Emsemble* das CNN com *XGBoost*.

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 bons resultados em quase todas as expressões, mas não obteve-se 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.

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	0.694901086653664
8	0.6285873502368348
9	0.6244079130677069
ensemble	0.7174700473669546

As matrizes de confusão estão normalizadas. Cada célula foi colorida de acordo com a incidência de elementos contidos nesta antes da normalização(quanto maior a incidência mais escura é a cor da célula). O *grid* de matrizes de confusões da Figura **??** foi gerado a partir da parte *PrivateTest* da base de dados, onde cada matriz corresponde a um modelo de CNN utilizado no *Ensemble*. As classes verdadeiras do exemplo estão representadas pelas linhas, enquanto, as colunas representam as classes preditas pelo *Ensemble*.

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.

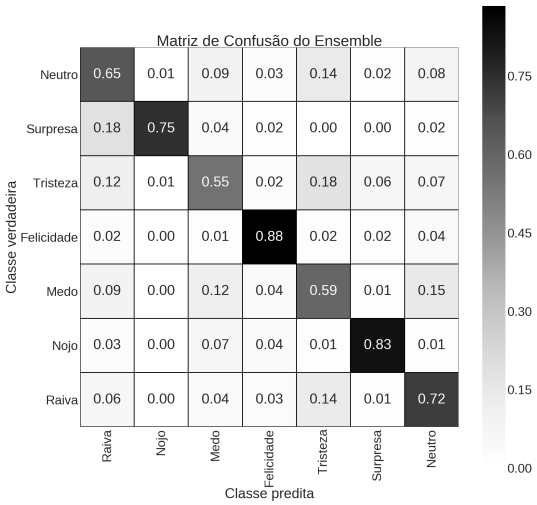


Figura 5: Matriz de Confusão do *Ensemble* (CNN + *XGBoost*)

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.

[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.

[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.

[Whitehill et al.] Whitehill, J., Bartlett, M. S., and Movellan, J. R. Automatic facial expression recognition. *Social emotions in nature and artifact*, 88.