Jorge Luis Melo Ribeiro

Meta-aprendizado aplicado ao Problema de Reconhecimento de Expressões Faciais

São Luís - MA

Setembro - 2018

Jorge Luis Melo Ribeiro

Meta-aprendizado aplicado ao Problema de Reconhecimento de Expressões Faciais

Projeto de Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, para aprovação no componente curricular Monografia I.

Curso de Ciência da Computação

UFMA

Orientador: Prof. Dr. Geraldo Braz Jr.

São Luís - MA

Setembro - 2018

Jorge Luis Melo Ribeiro

Meta-aprendizado aplicado ao Problema de Reconhecimento de Expressões Faciais

Projeto de Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, para aprovação no componente curricular Monografia I.

Projeto de Monografia Aprovado. Nota _____ (_____)

São Luís - MA, 02 de Setembro de 2018.

Prof. Dr. Geraldo Braz Jr.
Orientador

São Luís - MA
Setembro - 2018

Resumo

De acordo com o professor de psicologia Albert Mehrabian, estudioso da área de comunicação humana, 90% da expressão humana é não-verbal (MEHRABIAN et al., 1971). O homem, ao se comunicar, expressa de diversas formas aquilo que está sentido, especialmente através de reações. As reações faciais dão contexto e significado à fala humana, sendo os gestos executados de extrema importância para a compreensão do interlocutor. Diante disso, no contexto da evolução de sistemas inteligentes que tratam da interação com humanos, o entendimento correto da emoção sentida pelo homem pode auxiliar na resposta correta ou mais adequada retornada pelo sistema. Em muitos estudos e pesquisas se tem utilizado redes neurais para o treinamento e classificação de aplicações de aprendizado de máquina. Para se construir uma rede neural precisa-se primeiro escolher a sua arquitetura, sendo esta definida a partir de alguns parâmetros específicos, chamados de hiperparâmetros. As redes neurais convolucionais (CNNs) são utilizadas especificamente para problemas que envolvem imagens como entrada de dados, e também precisam ter seus hiperparâmetros definidos. A escolha e definição dos valores desses hiperparâmetros é um problema a ser resolvido nas redes neurais, pois precisa ser feito de forma empírica. Alguns otimizadores já são utilizados para realizar a escolha dos melhores hiperparâmetros, que fazem essa validação por tentativa e erro dentro de um espaço de busca. Alguns podem ser citados, como é o caso do Grid Search e do Random Search. Diante do exposto, a proposta deste trabalho é utilizar a biblioteca de otimização hyperopt para otimizar os hiperparâmetros de CNNs que irão realizar o treinamento e classificação de expressões faciais humanas.

Palavras-chaves: redes neurais, convolucionais, hiperparâmetros, otimização, hyperopt.

Sumário

1 Introdução

Expressões faciais possuem grande importância na comunicação entre humanos, pois elas dão contexto e complementam a expressão verbal. Logo, dentro da inteligência artificial tem se buscado reconhecer tais expressões dentro de 7 principais emoções: alegria, tristeza, nojo, medo, neutro, raiva e surpresa. O reconhecimento correto dessas expressões é um importante passo para a automatização de sistemas inteligentes. Muita pesquisa tem sido conduzida nos últimos anos, principalmente com o uso de redes neurais convolucionais (CNNs). A disponibilidade de datasets extensos e bem anotados, além de maior poder computacional contribuiu para o alcance de bons resultados utilizando CNNs.

Alguns datasets disponíveis como JAFFE (LYONS et al., 1998) ou Cohn Kanade (KANADE; TIAN; COHN, 2000) e Extended Cohn Kanade (LUCEY et al., 2010) foram gerados em laboratório sob condições controladas (rosto bem enquadrado, boa iluminação e resolução, expressões bem definidas) e nestes já se obtém resultados com acurácia acima de 90%. No entanto, reconhecer expressões faciais em condições naturais e diversas ainda é um desafio, pois todas as condições que podem ser aplicadas em laboratório podem não ocorrer em situações do dia a dia. Trabalhos recentes têm buscado utilizar variadas arquiteturas de CNNs na tarefa de realizar essa classificação, utilizando datasets com faces em condições naturais (chamados in-the-wild).

Algumas arquiteturas tornam-se conhecidas após obterem resultados de sucesso no ImageNet Challenge, desafio que ocorre uma vez por ano e disponibiliza um dos maiores datasets anotados, possuindo mais de 14 milhões de imagens (DENG et al., 2014). Nas últimas edições, as redes neurais convolucionais têm demonstrado sua grande capacidade de classificação de imagens, com a aplicação de CNNs cada vez mais profundas em cada edição.

Para se construir uma CNN, alguns aspectos importantes precisam ser levados em conta: dados que serão utilizados, tamanho do dataset, quantidade de camadas, tamanho de filtros das camadas convolucionais, porcentagem de dropout, entre outros. Logo, para cada problema, a escolha desses parâmetros deve ser bem definida para que se consiga obter um bom resultado. No contexto das CNNs, existem alguns parâmetros de suma importância, além dos já citados anteriormente, que são definidos antes do treinamento da rede. Tais parâmetros estão estritamente ligados à arquitetura da CNN, e por isso são chamados de hiperparâmetros. Após o grande crescimento das arquiteturas das CNNs em relação ao número de camadas, existe também a preocupação de se escolher os hiperparâmetros de forma eficaz. Isso pode ser feito por algumas técnicas já conhecidas, como o Grid Search e o Random Search (BERGSTRA et al., 2011). O Grid Search é uma técnica de força-bruta,

logo demanda muito tempo para ser executada. O Random Search é semelhante ao Grid Search, porém executa de forma estocástica e tem como proposta retornar os melhores hiperparâmetros em menos tempo quando comparado ao Grid Search.

Estudos recentes como (PINTO et al., 2009) e (COATES; NG, 2011) demonstram que o desafio de otimizar os hiperparâmetros em modelos profundos têm impedido o progresso científico. Por isso, seria adequado utilizar uma técnica em torno do processo de aprendizagem, de forma a realizar a escolha em um espaço de busca definido, que contenha intervalos específicos para cada hiperparâmetro. Nesse contexto, uma biblioteca chamada de hyperopt (BERGSTRA; YAMINS; COX, 2013) tem se popularizado. Nela, um espaço de busca é construído de forma estruturada e dois algoritmos de busca são fornecidos: Random Search e Tree-of-Parzen-Estimators (TPE) (BERGSTRA et al., 2011).

1.1 Justificativa

A compreensão automática por parte de um sistema inteligente da emoção sentida pelo ser humano em dado momento pode auxiliar na tomada de decisões corretas, dependendo da aplicação. Como mencionado na seção anterior, hoje existe a disponibilidade de uma grande variedade de datasets anotados de emoções em rostos humanos, e mais recente ainda, outros datasets in-the-wild extensos estão sendo disponibilizados pelos seus autores. Logo, a problemática que havia de ainda não existir datasets disponíveis de rostos em ações naturais pode ter sido superada. Em 2017, por exemplo, o dataset AffectNet (MOLLAHOSSEINI; HASANI; MAHOOR, 2017) que consta com mais de 1 milhão de imagens coletadas da internet está disponível para requisição de download.

Por isso, tendo os dados, pode-se utilizar de técnicas para treinamento e classificação, como é o caso das CNNs. Este trabalho ainda não utilizará de datasets extensos, e sim buscará metaotimizar os hiperparâmetros de redes neurais convolucionais por meio da biblioteca hyperopt, utilizando-se de datasets menores, como é o caso do JAFFE (LYONS et al., 1998), CK+ (LUCEY et al., 2010) e FER2013 (GOODFELLOW et al., 2013).

1.2 Objetivo

Com base no que foi introduzido e na justificativa, este trabalho tem como objetivo utilizar a biblioteca hyperopt de forma a estimar os melhores hiperparâmetros para arquiteturas de CNNs, para classificar emoções faciais em três datasets: JAFFE, CK+ e FER2013.

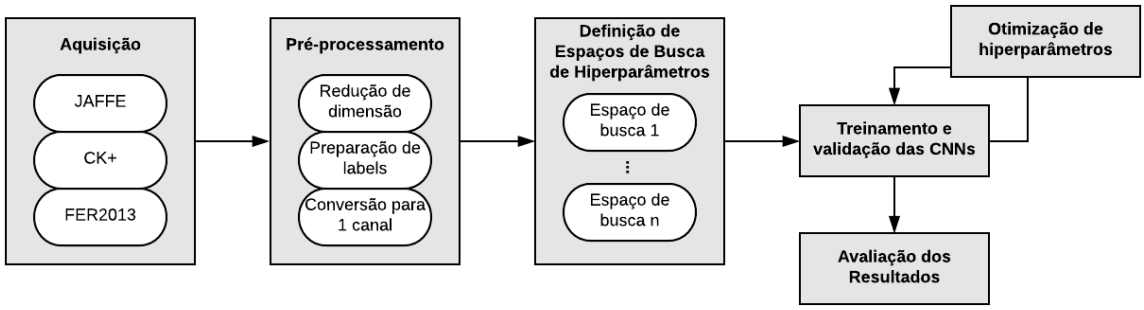
1.2.1 Objetivos Específicos

- Treinamento e classificação de emoções faciais por CNNs em sete categorias: alegria, tristeza, nojo, medo, neutro, raiva e surpresa;
- Utilização da biblioteca hyperopt de forma a otimizar a escolha dos hiperparâmetros das CNNs construídas;
- Verificar os hiperparâmetros que mais influenciam no treinamento das CNNs;
- Utilização de três diferentes datasets de forma a validar a metodologia;
- Buscar resultados semelhantes ou melhores ao estado-da-arte para trabalhos que utilizem os mesmos datasets.

2 Metodologia

Neste capítulo serão apresentadas as técnicas e métodos que serão utilizados para a realização deste trabalho. A metodologia consistirá em etapas bem definidas, que são: (1) Aquisição de datasets; (2) Pré-processamento dos datasets adquiridos; (3) Definição de espaço de busca de hiperparâmetros para cada arquitetura utilizada; (4) Treinamento, validação e otimização (execução) das CNNs. A Figura 1 representa as etapas citadas e descritas nas seções a seguir.

Figura 1 – Metodologia proposta.



Fonte: Autor

2.1 Aquisição de datasets

Os três datasets que serão utilizados nesse trabalho (JAFFE, CK+ e FER2013) serão reunidos e validados em suas anotações. Os três estão disponíveis publicamente na internet. O dataset JAFFE (Japanese Female Facial Expression) consta com 213 imagens de 10 pessoas de etnia japonesa do sexo feminino. O dataset CK+ (Extended Cohn Kanade), que é um dataset aumentado do CK (KANADE; TIAN; COHN, 2000), consta de 593 sequências de 123 pessoas. A última imagem (expression peak) de cada sequência será utilizada para realizar o treinamento com a emoção determinada pela sua anotação. Por último o FER2013, dataset disponível a partir de um challenge do Kaggle (KAGGLE. . . , 2018) do ano de 2013. Contém mais de 30 mil imagens de rostos in-the-wild e já foi utilizado em trabalhos anteriores. Os tamanhos e formatos de cada dataset diferem um do outro, por isso as arquiteturas que serão construídas obedecerão essas diferenças. Exemplos de imagens dos três datasets citados podem ser vistos nas Figuras 2 e 3.

Figura 2 – Exemplos de imagens dos datasets CK+ (acima) e JAFFE (abaixo).

Fonte: Adaptado de (PRAMERDORFER; KAMPEL, 2016)

Figura 3 – Exemplos de imagens do dataset FER2013.

Fonte: (HOLDER; TAPAMO, 2017)

2.2 Pré-processamento

Algumas etapas serão realizadas antes da utilização de cada dataset em CNNs. No dataset JAFFE, as imagens possuem dimensão 256x256 pixels e serão reduzidas para agilizar o processo de treinamento. Também será necessário gerar um arquivo de anotações, pois o dataset não possui um. Cada imagem indica em seu nome de arquivo qual emoção está representada.

Por exemplo, na Figura 4 é possível visualizar que a sigla após o primeiro ponto no nome de cada arquivo indica a emoção realizada pelo sujeito. FE indica emoção medo, HA indica emoção alegria, SU indica emoção surpresa, e assim por diante. Com isso, o arquivo de anotações será gerado a partir da leitura do nome de cada imagem pertencente ao dataset.

O dataset CK+ disponibiliza as anotações prontas para o pico de expressão em cada sequência de frames. Será necessário apenas reunir cada anotação em um arquivo comum, de forma a facilitar a leitura. As imagens do dataset possuem três canais, por isso serão convertidas para um canal e terão sua dimensão reduzida, também para agilizar o treinamento.

Figura 4 – Visualização dos nomes das imagens presentes no dataset JAFFE.

Fonte: Autor

O dataset FER 2013 já foi previamente pré-processado, utilizando-se de um algoritmo de detecção da faces Haar Cascade da biblioteca OpenCV, de forma a eliminar imagens de rostos mal enquadrados ou difíceis de serem detectados. Ainda será realizada uma redução de tamanho do dataset, para apenas metade das imagens serem utilizadas (pouco mais de 10 mil imagens serão mantidas).

2.3 Definição de espaço de busca de hiperparâmetros

O espaço de busca de hiperparâmetros precisa ser construído obedecendo a estrutura do hyperopt. Essa tarefa pode ser feita de várias formas, porém a forma mais comum é instanciar um dict em Python, nomeando cada hiperparâmetro e definindo seu intervalo de busca.

Um espaço de busca distinto precisará ser construído para cada arquitetura utilizada, devido ao fato das arquiteturas possuírem estrutura própria. Logo, as arquiteturas escolhidas precisam ser bem compreendidas, de forma a selecionar os melhores hiperparâmetros que irão compor o espaço de busca e seus respectivos intervalos. Um exemplo de espaço de busca pode ser visto na Figura 5.

Figura 5 – Exemplo de um espaço de busca definido para uma Multilayer Perceptron.

```
space = {'layer_size':hp.quniform('layer_size', 25, 100, 1),
         'alpha':hp.lognormal('alpha', mu=np.log(1e-4), sigma=1),
         'solver':hp.choice('solver', ['lbfgs', 'sgd', 'adam']),
         'activation':hp.choice('activation', ['logistic', 'tanh', 'relu']),
         'learning_rate':hp.loguniform('learning_rate', low=np.log(1e-4), high=np.log(1.)),
        }
```

Fonte: Autor

2.4 Treinamento, validação e otimização dos hiperparâmetros das CNNs

Na última etapa serão realizados os testes propriamente ditos, de forma a validar a metodologia. As CNNs serão construídas com a API Keras, que será executada no topo da biblioteca de aprendizado de máquina TensorFlow. As execuções serão feitas em GPU, para se obter execuções mais rápidas. Em torno das execuções, a biblioteca hyperopt irá realizar as otimizações adequadas dos hiperparâmetros, a partir do loss obtido em cada execução.

3 Resultados Esperados

Neste trabalho espera-se validar a metaotimização dos hiperparâmetros de CNNs na tarefa de classificar emoções em rostos humanos. A validação será concluída ao se observar resultados equiparáveis aos que já existem na literatura, para se comprovar que a escolha dos hiperparâmetros pode ser uma tarefa automatizada.

Busca-se também fornecer avanços na tarefa de otimização de hiperparâmetros e resultados melhores no desafio de classificar emoções humanas.

4 Cronograma

Nesta seção serão numeradas as atividades a serem realizadas para a conclusão desse trabalho. Na tabela abaixo é possível visualizar a divisão de cada atividade por período de tempo.

a) Levantamento bibliográfico

b) Aquisição e pré-processamento de datasets

c) Seleção de arquiteturas de CNNs que serão utilizadas

d) Implementação das arquiteturas selecionadas em Keras

e) Seleção dos espaços de busca de hiperparâmetros para cada arquitetura selecionada

f) Verificação dos hiperparâmetros escolhidos

g) Treinamento, validação e otimização das CNNs

h) Escrita da monografia

i) Defesa da monografia

Tabela 1 – Cronograma de Atividades

	Agosto	Setembro	Outubro	Novembro	Dezembro
a)	x				
b)	x				
c)	x	x			
d)		x			
e)		x			
f)		x			
g)		x	x	x	
h)				x	x
i)					x

Referências

BERGSTRA, J.; YAMINS, D.; COX, D. D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In: CITESEER. *Proceedings of the 12th Python in Science Conference*. [S.l.], 2013. p. 13–20. Citado na página 6.

BERGSTRA, J. S. et al. Algorithms for hyper-parameter optimization. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2011. p. 2546–2554. Citado 2 vezes nas páginas 5 e 6.

COATES, A.; NG, A. Y. The importance of encoding versus training with sparse coding and vector quantization. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. [S.l.: s.n.], 2011. p. 921–928. Citado na página 6.

DENG, J. et al. Scalable multi-label annotation. In: ACM. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. [S.l.], 2014. p. 3099–3102. Citado na página 5.

GOODFELLOW, I. J. et al. Challenges in representation learning: A report on three machine learning contests. In: SPRINGER. *International Conference on Neural Information Processing*. [S.l.], 2013. p. 117–124. Citado na página 6.

HOLDER, R. P.; TAPAMO, J. R. Improved gradient local ternary patterns for facial expression recognition. *EURASIP Journal on Image and Video Processing*, Springer, v. 2017, n. 1, p. 42, 2017. Citado na página 9.

KAGGLE webpage. 2018. <https://www.kaggle.com/>. Citado na página 8.

KANADE, T.; TIAN, Y.; COHN, J. F. Comprehensive database for facial expression analysis. In: IEEE. *fg*. [S.l.], 2000. p. 46. Citado 2 vezes nas páginas 5 e 8.

LUCEY, P. et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. [S.l.], 2010. p. 94–101. Citado 2 vezes nas páginas 5 e 6.

LYONS, M. et al. Coding facial expressions with gabor wavelets. In: IEEE. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. [S.l.], 1998. p. 200–205. Citado 2 vezes nas páginas 5 e 6.

MEHRABIAN, A. et al. *Silent messages*. [S.l.]: Wadsworth Belmont, CA, 1971. v. 8. Citado na página 3.

MOLLAHOSSEINI, A.; HASANI, B.; MAHOOR, M. H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017. Citado na página 6.

PINTO, N. et al. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, Public Library of Science, v. 5, n. 11, p. e1000579, 2009. Citado na página 6.

PRAMERDORFER, C.; KAMPEL, M. Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*, 2016. Citado na página 9.