Projeto de Pesquisa

Iniciação Científica

Deep Learning aplicado ao reconhecimento facial

Estudante: Kaique Nunes de Oliveira
Bacharelado em Ciência da Computação

Orientadora: Nina S. T. Hirata
Departamento de Ciência da Computação

Instituto de Matemática e Estatística
Universidade de São Paulo

Resumo

Historicamente o reconhecimento facial por computadores frequentemente foi visto como algo de ficção científica. Porém, nos últimos anos, foram desenvolvidos métodos que são capazes de realizar o reconhecimento de face com taxa de acerto bastante elevada, fazendo com que essa tecnologia esteja cada vez mais presente no dia a dia das pessoas. Esse grande avanço ocorreu principalmente devido às contribuições trazidas pelo *Deep Learning*, a disponibilização de vastos bancos de dados públicos de fotos de rostos humanos, e às Redes Neurais Convolucionais (CNN). CNNs realizam a extração de características de imagens e aprendem representações de alto nível a partir dos dados de treinamento, sendo muito utilizada, para além de reconhecimento facial, em várias tarefas na área de Visão Computacional. Este projeto de pesquisa visa o desenvolvimento de um modelo para reconhecimento de faces, inicialmente utilizando datasets publicamente disponíveis. Neste processo, o estudante terá a oportunidade de adquirir conhecimentos nas áreas de Visão Computacional, Aprendizado de Máquina e principalmente de *Deep Learning*, além do processo de investigação científica. Também terá a oportunidade de desenvolver habilidades práticas para a implementação e treinamento de CNNs. Ao final, espera-se que seja produzido um aplicativo para testar esses modelos com imagens de faces coletadas no ambiente universitário.

São Paulo, 28 de setembro de 2023

1 Introdução

Nos últimos anos houve uma grande evolução nos sistemas de reconhecimento facial. O uso dessa tecnologia se torna cada vez mais presente na vida pessoal dos cidadãos, no setor público e também no privado. Um dos usos mais comuns dessa tecnologia é na área da segurança e autenticação de usuários; por exemplo, no controle de entrada de pessoas em determinados ambientes, ou no desbloqueio de uma simples tela de celular.

Tendo isso em vista, o grande salto qualitativo mais recente das técnicas de reconhecimento facial ocorreu devido ao emprego de técnicas de *Machine Learning*, mais especificamente de *Deep Learning*, como pode ser visto na ilustração da figura 1. *Deep Learning* (Goodfellow et al., 2016) é uma subárea da Inteligência Artificial (IA) e de *Machine Learning* (ML) que utiliza redes neurais com múltiplas camadas para processar os dados e automaticamente extrair deles as características mais eficazes para a inferência final.

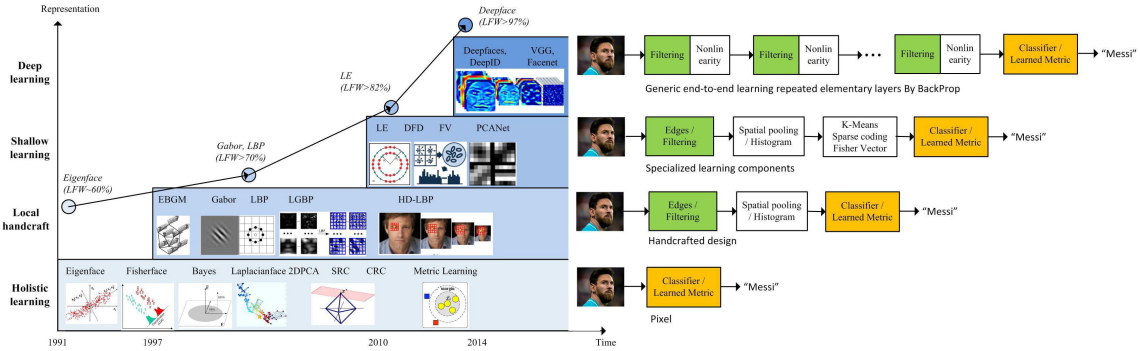

Figura 1: Imagem com um gráfico que mostra os saltos da capacidade de acerto dos diferentes modelos de reconhecimento facial e uma representação de como os grupos de modelo funcionam. Observe como, a partir de mais ou menos 2014, o emprego de técnicas de *Deep Learning* gerou modelos com acerto acima de 90% quando aplicados ao dataset LFW (Huang et al., 2007). (Fonte: Wang and Deng (2021))

Os trabalhos iniciais que tratavam de *deep learning* aplicado ao reconhecimento facial e que adquiriram grande notoriedade geralmente utilizavam datasets privados para o treinamento das redes. Um exemplo foi o 'DeepFace' (Taigman et al., 2014), do time de IA do Facebook. O DeepFace utilizou mais de 4 milhões de imagens representantes de mais de 4 mil pessoas diferentes. Nesse sentido, esses grandes trabalhos eram difíceis de serem replicados pela comunidade científica na medida em que eles utilizavam datasets privados. Para solucionar este problema, na última década surgiram diversos datasets gigantes e que todos podemos utilizar, o que permite um treinamento mais poderoso e abrangente de modelos de aprendizado de máquina criados por diferentes grupos. Na tabela 2, podemos ver uma compilação de datasets públicos e privados existentes em 2021 para a tarefa de re-

conhecimento facial. O 'DeepFace' e os trabalhos recentes baseados em técnicas de *Deep Learning* e que tratam de reconhecimento facial tipicamente empregam Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*).

The commonly used FR datasets for training

Datasets	Publish Time	#photos	#subjects	# of photos per subject [1]	Key Features
MS-Celeb-1 M (Challenge 1)[45]	2016	10 M 3.8 M(clean)	100,000 85 K(clean)	100	breadth; central part of long tail; celebrity; knowledge base
MS-Celeb-1 M (Challenge 2)[45]	2016	1.5 M(base set) 1 K (novel set)	20 K(base set) 1 K(novel set)	1/-/100	low-shot learning; tailed data; celebrity
MS-Celeb-1 M (Challenge 3) [163]	2018	4 M(MSv1c) 2.8 M (Asian-Celeb)	80 K(MSv1c) 100 K (Asian-Celeb)	-	breadth;central part of long tail; celebrity
MegaFace [44,164]	2016	4.7 M	672,057	3/7/2469	breadth; the whole long tail;commonalty
VGGFace2 [39]	2017	3.31 M	9,131	87/362.6/843	depth; head part of long tail; cross pose, age and ethnicity; celebrity
CASIA WebFace [120]	2014	494,414	10,575	2/46.8/804	celebrity
MillionCelebs [165]	2020	18.8 M	636 K	29.5	celebrity
IMDB-Face [124]	2018	1.7 M	59 K	28.8	celebrity
UMDFaces-Videos [166]	2017	22,075	3,107	–	video
VGGFace [37]	2015	2.6 M	2,622	1,000	depth; celebrity; annotation with bounding boxesand coarse pose
CelebFaces+ [21]	2014	202,599	10,177	19.9	private
Google [38]	2015	>500 M	>10 M	50	private
Facebook [20]	2014	4.4 M	4 K	800/1100/1200	private

Figura 2: Alguns dos datasets públicos e privados que foram criados na última década. (Fonte: Wang and Deng (2021))

O problema de reconhecimento facial, por ser um importante problema com diversas possibilidades de aplicação prática, e pelo fato de ser um problema já bastante investigado e para o qual existem datasets públicos com diferentes características, é bastante adequado como objeto de estudo para introduzir um estudante a diferentes ramos do conhecimento. Em particular, trabalhar este problema envolve conceitos relacionados a processamento de imagens (Gonzalez and Woods, 2002), visão computacional (Prince, 2012; Szeliski, 2011), *machine learning* (Abu-Mostafa et al., 2012) e *deep learning* (Goodfellow et al., 2016; Nielsen, 2015), além de outros fundamentos e conceitos subjacentes a esses ramos do conhecimento.

2 Objetivos

O principal objetivo deste projeto de pesquisa é o desenvolvimento de um modelo para reconhecimento facial. Para tanto, definimos os seguintes objetivos específicos:

- Estudo de conceitos e fundamentos importantes para compreender o problema (reconhecimento de faces a partir de imagens de faces) e para compreender os modelos baseados em *deep learning*, utilizados nas soluções modernas
- Adquirir habilidades de programação importantes para o treinamento e avaliação de redes neurais
- Revisar a literatura da área (reconhecimento de faces) e selecionar datasets e arquiteturas de CNNs para serem exploradas experimentalmente

- Planejar os experimentos, executá-los e avaliar os resultados

Como explicado, o problema de reconhecimento facial (RF) é um problema clássico de visão computacional. Como podemos ver em (Guo and Zhang, 2019), quando um sistema de RF recebe uma imagem, quatro passos se seguem: detecção da face, alinhamento da face, extração das features da face, e, finalmente, vem o "face matching". O face matching pode ser pensado para atacar dois tipos de problemas: Verificação Facial (VF) ou Identificação Facial (IF). O primeiro, VF, diz, ao receber duas imagens de rostos, se elas são da mesma pessoa. O segundo, IF, diz, ao receber uma imagem de um rosto, qual a identidade da pessoa da imagem dado um conjunto de identidades previamente listadas. Nesse sentido, um objetivo secundário é o desenvolvimento de um aplicativo, associado a um pequeno banco de imagens de faces a serem coletadas no ambiente universitário, que seja capaz de capturar a imagem de uma face e identificar a pessoa se a mesma estiver registrada no banco ou então indicar que não foi possível reconhecer a pessoa, ou seja, um aplicativo que lide com o problema de IF.

O desenvolvimento do projeto de pesquisa, além de complementar e aprofundar a formação do estudante em importantes áreas do conhecimento, propiciará exposição e treinamento em processos de investigação científica, que deverá culminar na elaboração de um relatório científico final.

3 Material e métodos

O desenvolvimento do projeto deverá seguir de forma bastante próxima os objetivos específicos listados acima. Desta forma, organizamos esta seção de forma a detalhar os materiais e métodos a serem empregados para cada um dos objetivos específicos acima.

3.1 Estudo de conceitos e fundamentos

O problema a ser tratado neste projeto é um típico problema de Visão Computacional. Desta forma, em termos de conceitos e fundamentos teóricos é desejável que o estudante possua conhecimentos sobre processamento de imagens (Gonzalez and Woods, 2002) e aprendizado de máquina (Abu-Mostafa et al., 2012). Para estudar e implementar CNNs, é necessário também um conhecimento sólido de redes neurais (Nielsen, 2015).

Neste sentido, é importante ressaltar que o estudante participa atualmente de um grupo de estudos sobre aprendizado de máquina, formado por alunos de graduação. Esse grupo reúne-se semanalmente para estudar e discutir aspectos teóricos

e práticos de aprendizado de máquina. Neste grupo, o estudante está tendo a oportunidade de aprender redes neurais e aplicações em problemas de Visão Computacional e de Processamento de Linguagem Natural.

Desta forma, prevê-se que em pouco tempo o estudante já terá reunido conhecimentos suficientes para o estudo de redes neurais convolucionais (CNNs).

Redes neurais convolucionais (CNNs): São redes neurais que possuem camadas formadas por nós (unidades de processamento) que implementam os filtros de convolução (Goodfellow et al., 2016). Esses filtros são parametrizados por kernels, utilizados para realizar processamentos locais ao longo de toda a extensão da imagem. Os kernels, implementados como matrizes de pesos, tem os valores estabelecidos no processo de treinamento da rede. Isto significa que os filtros são "aprendidos" pela rede durante o seu treinamento de forma otimizada para a tarefa-alvo do treinamento.

Alguns possíveis materiais para o estudo de CNNs estão listados a seguir, porém outros deverão ser utilizados à medida que o estudante for ganhando familiarização com o tópico.

- http://cs231n.stanford.edu/, material disponibilizado pela Universidade de Stanford
- Artigo *Convolutional Networks and Applications in Vision* (LeCun et al., 2010)
- Livro *Deep Learning* (Goodfellow et al., 2016)
- Artigos bem conhecidos sobre modelos CNN para classificação de imagem (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016)

Outros conceitos: Adicionalmente, deverá ser dada atenção à formulação do problema de reconhecimento de faces. Em princípio, pretendemos considerar a situação em que as faces a serem reconhecidas já estão recortadas devidamente. As CNNs deverão ser treinadas para gerar codificações únicas (ou muito similares) para diferentes fotos de um mesmo indivíduo. Em um primeiro momento esse problema pode ser abordado como um problema de classificação no qual cada indivíduo corresponde a uma classe. Posteriormente, deverão ser estudadas as arquiteturas e estratégias utilizadas para gerar codificações (*embeddings*) distintas para instâncias de classes distintas (por exemplo, as técnicas de *contrastive learning* (Khosla et al., 2020; Jaiswal et al., 2020)). Neste caso, o problema deve ser formulado como um problema de busca da imagem mais similar em um banco de dados, denominado

acima como Identificação Facial (IF).

3.2 Aquisição de habilidades de programação em *deep learning*

Apesar de o estudante ser do curso de Bacharelado em Ciência da Computação, a área de *deep learning* requer familiarização com as bibliotecas comumente utilizadas (como PyTorch ou Keras/TensorFlow), além de bibliotecas do Python como `numpy`, `pandas`, ou de visualização.

Um treinamento preliminar para adquirir familiaridade com essas bibliotecas deverá ser realizado por meio de preparo e execução de *scripts* para treinamento e avaliação de modelos treinados sobre problemas de classificação mais simples (por exemplo, MNIST ou CIFAR-10). Prevê-se que essa atividade ocorrerá naturalmente durantes as atividades junto ao grupo de estudos do qual o estudante participa.

3.3 Revisão de literatura

O estudante já realizou uma revisão preliminar da literatura da área, para familiarizar-se e entender minimamente o problema de reconhecimento de faces e quais tipos de resultados estão sendo alcançados.

Essa revisão deverá ser aprofundada durante o desenvolvimento do projeto, visando identificar as principais abordagens utilizadas nos trabalhos publicados. Especificamente, é do interesse identificarmos quais formulações de problema são considerados, quais datasets são utilizados, quais tipos de arquiteturas são utilizados, e quais métricas de avaliação são consideradas nesses trabalhos.

Na revisão preliminar realizada, observamos que a fim de gerar comparações de performance entre redes neurais, muitas vezes elas são aplicadas também ao LFW (*Labeled Faces in the Wild* (Huang et al., 2007)), um dataset público e relativamente pequeno. Ao longo dos anos novas versões do LFW foram lançadas e atendem diferentes critérios. Esse dataset é interessante pois ele traz imagens de rostos em ambientes comuns, fora de um lugar controlado como um laboratório, por exemplo. Essa característica permite treinar e avaliar a performance de redes neurais em um ambiente mais realista, ainda que com suas limitações. Esse dataset, devido ao seu uso frequente nos trabalhos, e devido ao seu tamanho reduzido, será utilizado pelo estudante durante o avanço do projeto.

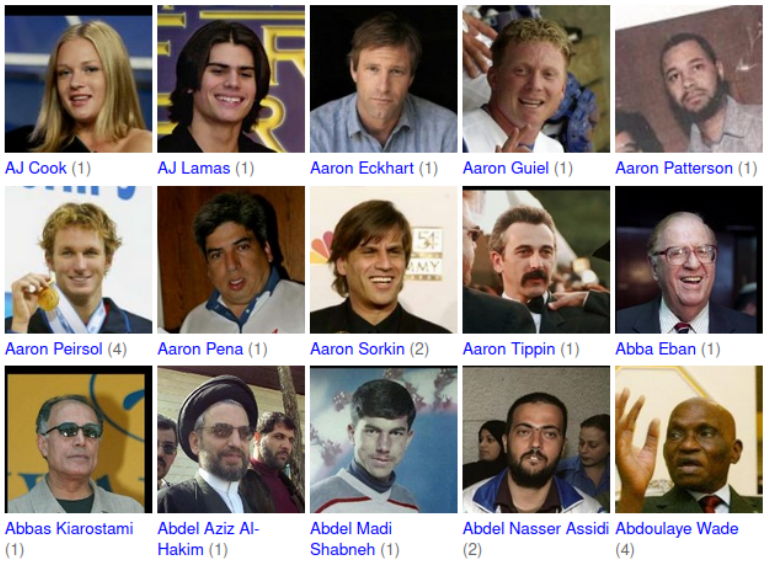

Figura 3: Exemplos de algumas imagens do LFW. (Fonte: http://vis-www.cs.umass.edu/lfw/#information)

3.4 Parte experimental

Uma vez estudados os conceitos e fundamentos, feita a familiarização com programação no que diz respeito ao treinamento e avaliação de CNNs, e identificados os datasets, arquiteturas e métricas de avaliação geralmente utilizados, será trabalhado o planejamento e execução dos experimentos computacionais.

Um primeiro passo será a seleção de datasets e arquiteturas a serem utilizados e métricas de avaliação a serem empregadas. O segundo passo será o planejamento dos experimentos. Nesta parte deverão ser abordadas questões tais como o uso ou não de *transfer learning* (emprego de redes pré-treinadas em outro conjunto de dados como ponto de partida) e a configuração dos diversos hiperparâmetros importantes no treinamento da rede (como taxa de aprendizado, tamanho do *batch* de treinamento, função de perda, uso ou não de regularização, entre outros). Uma questão importante diz respeito à escolha de um modelo final, caso diferentes arquiteturas ou diferentes hiperparâmetros sejam empregados. Para a comparação de modelos, as métricas identificadas durante o estudo da literatura da área serão utilizadas.

Outra parte importante será a análise de resultados. Além de métricas objetivas que quantificam o desempenho global, deveremos explorar também técnicas que permitem uma avaliação mais qualitativa. Por exemplo, uma dessas técnicas é o mapa de projeção gerado pelo algoritmo t-SNE (van der Maaten and Hinton, 2008), que permite avaliar se os *embeddings* calculados pelas CNNs estão de fato agrupando adequadamente as instâncias de uma mesma classe. Outras avaliações mais qualitativas podem também ser realizadas por meio de inspeção visual das imagens, principalmente as que não forem reconhecidas corretamente, em busca de

algum padrão visual que possa explicar os erros.

Todos os detalhes dos experimentos realizados, assim como das análises realizadas, deverão ser documentados. Os códigos desenvolvidos serão disponibilizados no `github`.

3.5 Outras informações

Uma parte importante da iniciação científica é a introdução do estudante à pesquisa científica. Neste sentido, além de já estar participando e acompanhando a elaboração deste projeto, o estudante será exposto a vários aspectos da investigação científica durante a execução deste projeto.

Pretendemos também desenvolver um aplicativo para testar algum modelo de reconhecimento de faces em um cenário real. Especificamente, pretendemos coletar imagens de faces (com o consentimento das pessoas envolvidas) no ambiente universitário e criar um pequeno banco de dados, e dada uma imagem de face qualquer, o aplicativo deverá responder se ela corresponde a uma das pessoas cadastradas no banco ou não. Neste contexto, dependendo do andamento, poderão ser explorados também imagens de vídeo e modelos para detecção de faces, a ser empregado em uma etapa anterior ao reconhecimento.

Com respeito à comunicação científica, caso pertinente, pretendemos trabalhar a elaboração de artigos para serem submetidos a eventos de divulgação científica tais como o SIICUSP (Simpósio Internacional de Iniciação Científica e Tecnológica da USP) ou o WUW/SIBGRAPI, em momentos oportunos.

4 Plano de trabalho e cronograma de execução

As atividades a serem desenvolvidas estão descritas a seguir.

1. **Estudo de fundamentos:** Corresponde às atividades descritas na seção 3.1
2. **Prática em programação com *deep learning*:** Corresponde às atividades descritas na seção 3.2
3. **Revisão de literatura** Corresponde às atividades descritas na seção 3.3
4. **Parte experimental:** De acordo com o descrito na seção 3.4 compreende as seguintes atividades:
   (a) Seleção de datasets e arquiteturas a serem usadas

(b) Planejamento dos experimentos

(c) Execução dos experimentos

(d) Registro dos experimentos e resultados

(e) Análise de resultados

5. **Desenvolvimento de um aplicativo:** para testar como esses modelos funcionam com dados coletados no ambiente universitário

6. **Treinamento em escrita científica:** Como parte das atividades estão previstas o treinamento de escrita científica. Este será por meio da elaboração de relatório científico final. Eventualmente, dependendo dos resultados alcançados, poderá também ser elaborado um artigo científico a ser submetido a um fórum adequado.

Um cronograma aproximado para a execução dessas atividades está apresentado a seguir.

	Meses											
Atividade	1	2	3	4	5	6	7	8	9	10	11	12
1	x	x	x									
2		x	x	x								
3	x	x	x	x	x	x	x	x				
4a			x	x								
4b				x	x							
4c					x	x	x	x	x	x		
4d						x	x	x	x	x		
4e						x	x	x	x	x		
5									x	x	x	
6										x	x	x

5 Forma de análise dos resultados

Os resultados da pesquisa serão avaliados em termos dos seguintes produtos ou atuações resultantes como decorrência direta da execução do projeto de pesquisa:

- Disponibilização pública dos códigos desenvolvidos.

- Submissão ou publicação do trabalho em eventos ou outros veículos pertinentes.

- Apresentação dos resultados em congressos e eventos científicos locais e regionais.

- Divulgação dos conhecimentos adquiridos através de atividades no ambiente universitário, como atividades em grupos de extensão ou palestras.

Referências

Abu-Mostafa, Y. S., Lin, H.-T., and Magdon-Ismail, M. (2012). *Learning From Data*. AMLBook.

Gonzalez, R. C. and Woods, R. E. (2002). *Digital Image Processing*. Addison-Wesley Publishing Company, second edition.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Guo, G. and Zhang, N. (2019). A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256.

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press.

Prince, S. D. J. (2012). *Computer Vision – Models, Learning and Inference*. Cambridge.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Szeliski, R. (2011). *Computer Vision – Algorithms and Applications*. Springer.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Wang, M. and Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, 429:215–244.