Redes Convolucionais para reconhecimento facial através de imagens de regiões oculares

Eliana Pereira da Silva, José Hiroki Saito
Centro Universitário Campo Limpo Paulista – UNIFACCAMP
Campo Limpo Paulista – SP – Brasil

eliana.pereiras@gmail.com, saito@cc.faccamp.br

***Abstract**: This paper describes face recognition using partial images containing only the ocular region. For this purpose, three convolutional networks are considered: Neocognitron, LeNet and AlexNet. During the preliminary phase of the experiments, the pre-processed face image database of Essex University, containing 395 images between men and women faces, with 20 images per individual, was used. The preliminary results indicated that, although the face recognition are feasible, the ocular region image cropping morphology should be modified for more accurate classification, and larger database should be used for the training and test phases.*

***Resumo**: O presente artigo descreve o reconhecimento de faces utilizando imagens parciais, contendo somente a região ocular. Para esse propósito são usadas três redes convolucionais: Neocognitron, LeNet e AlexNet. Durante a fase preliminar dos experimentos, foi utilizada a base de dados de imagens faciais, pré-processada da Universidade Essex, com 395 imagens faciais entre homens e mulheres, sendo 20 imagens por indivíduo. Os resultados preliminares indicaram que, embora o reconhecimento de face seja viável, a forma da captura das imagens de regiões oculares deveria ser modificada para melhor classificação, e uma base de dados maior deveria ser usada para as fases de treinamento e teste.*

1 Introdução

As redes neurais convolucionais têm se apresentado como uma boa alternativa nos diferentes problemas, como reconhecimento de caracteres (Fukushima, 1980), (LeCun et al., 1998), (Bengio e LeCun, 2007), reconhecimento facial (Saito et al., 2005). O presente artigo pretende investigar de forma comparativa o uso das redes neurais convolucionais Neocognitron, LeNet-5 e AlexNet, no reconhecimento facial utilizando como padrão de entrada, imagens da região ocular da face. A região ocular é altamente discriminatória devido à grande diferença de detalhes da expressão ocular entre os indivíduos. Este texto, além desta seção introdutória, na Seção 2 são descritas as redes neurais utilizadas neste estudo; na Seção 3 é descrito o banco de imagens utilizado; na Seção 4 são apresentados os experimentos realizados; e na Seção 5 são descritos: a análise dos resultados, as considerações finais e trabalhos futuros.

2 Redes neurais convolucionais

As redes neurais convolucionais se caracterizam pelas conexões em pequenas regiões, denominados campos receptivos, dos seus neurônios. Esses neurônios se organizam matricialmente, acompanhando a camada de entrada, e o seu processamento resulta numa matriz denominada matriz de características. Essa matriz de características passa a ser a camada de entrada para os neurônios da camada seguinte. Cada matriz de características é denominado de plano celular. Num plano celular todos os seus neurônios tem a mesma característica, e o

processamento conjunto desses neurônios funciona como uma aplicação de filtros convolucionais ao longo de todas as posições da sua entrada, detectando ou filtrando uma determinada característica ao longo de todas as posições. Numa determinada camada são dispostos vários planos celulares, de diferentes tipos de filtros. O conjunto de todos os planos de uma camada contém todas as características relevantes da imagem de entrada, correspondente àquele nível hierárquico. Uma outra característica dessas redes é a existência de camadas de redução do tamanho do mapa de características, denotadas *pooling* que implica em selecionar dentre um campo receptivo uma das saídas. Assim, se o campo receptivo for de quatro neurônios, implica na redução para um quarto, caso a camada de *pooling* não sobreponha as entradas. O conjunto de uma camada convolucional e uma camada de *pooling* pode ser denominado de estágio. Após uma sequência de estágios, seguem as camadas totalmente conectadas. Essas camadas servem para coletar as informações existentes em vários planos celulares da sua camada anterior para o reconhecimento de uma determinada classe ou padrão. Cada plano celular das camadas totalmente conectadas representa uma determinada classe.

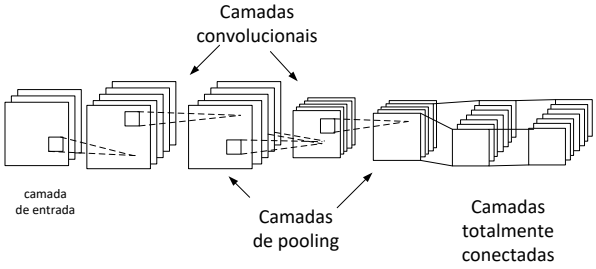

Figura 1. Um esboço de uma rede neural convolucional.

Na Figura 1 é mostrado um diagrama de uma rede convolucional, onde na extrema esquerda se encontra a camada de entrada. Logo ao seu lado direito se encontra a primeira camada convolucional, com diversos planos celulares, e em seguida a primeira camada de *pooling*. A rede termina com as camadas totalmente conectadas. Nota-se que a estrutura de uma rede convolucional é hierárquica, sendo que as camadas à direita da Figura 1 são hierarquicamente superiores em relação às camadas da esquerda.

2.1 Neocognitron

A rede convolucional Neocognitron foi proposta por Fukushima para reconhecimento de dígitos manuscritos (Fukushima, 1980). No Neocognitron a camada convolucional é denotado de camada U_S, de células-S responsáveis pela extração de características, e a camada de *pooling* é denotado de camada U_C, de células-C que aplica o filtro da média. As características mais simples como bordas e linhas em direções variadas são extraídas pelos estágios iniciais. Nos estágios seguintes são reconhecidos os fatores mais complexos como ângulos, extremidades e polígonos. A figura 2 é uma ilustração do processo de reconhecimento da letra 'A'.

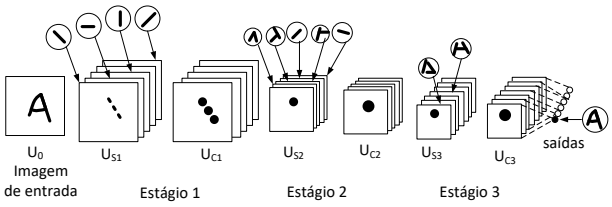

Figura 2- Fatores reconhecidos pelos neurônios nos 3 estágios da rede.

Uma célula-S é calculada usando a equação 1, onde, no numerador, o primeiro somatório representa todos os planos celulares conectados, o segundo somatório representa todas as conexões dentro do campo receptivo num plano de células-C; e cada entrada de uma célula-C, u_c é multiplicada pelo respectivo peso *a*. No denominador, Θ é o limiar responsável pela habilidade de extrair características, *b*

representa um peso, e v_c é um valor correspondente a uma célula auxiliar denotada célula-V.

$$u_s = \varphi[x] = \varphi\left[\frac{1+\sum_{planos}\sum_{recep} a \times u_c}{1+\theta \times b \times v_c} - 1\right] \quad (1)$$

No resultado do cálculo do argumento *x* da equação 1 é aplicada a função de ativação φ, definida pela equação 2, também conhecida como ReLU (*Rectified Linear Unit*):

$$\varphi[x] = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} \quad (2)$$

Nota-se que o argumento *x* deve ser maior ou igual a zero quando o numerador $1 + \sum_{planos}\sum_{recep} a \times u_c$ for maior ou igual ao denominador $1 + \theta \times b \times v_c$. Se o denominador for maior que o numerador, evita o disparo da célula u_s, a função do denominador é inibitória. Os pesos *a* e *b* são variáveis e modificados durante o treinamento. Em geral esses pesos são maiores ou iguais a zero.

Uma célula-V é calculada como raiz quadrada da soma ponderada pelo peso *c*, do quadrado de entradas u_c, conforme equação 3. O peso *c* é fixo, sendo o valor normalizado maior no centro do campo receptivo, decrescendo monotonicamente em direção radial.

$$v_c = \sqrt{\sum_{planos}\sum_{recep} c \times u_c^2} \quad (3)$$

A equação 4 é relativa à célula-C, denotado u_c, calculado pelo somatório do campo receptivo das células u_s ponderadas pelo peso *d*, que como o peso *c*, decresce monotonicamente em direção radial, porém, os seus valores não são normalizados.

$$u_c = \Psi[x] = \Psi\left[\sum_{recep} d \times u_s\right] \quad (4)$$

A função $\boldsymbol{\Psi}$ é descrita pela equação 5, que é uma forma de normalização do argumento *x*.

$$\Psi[x] = \begin{cases} \frac{x}{1+x} \; se \; x \geq 0 \\ 0, \; caso \; contrário \end{cases} \quad (5)$$

2.1.1 Treinamento da rede Neocognitron

Conforme visto, os pesos *a* e *b* são ajustados e os pesos *c* e *d* são fixos, portanto, os neurônios das camadas-S, que manipulam os pesos *a* e *b*, precisam de treinamento. O aprendizado é competitivo, não-supervisionado, e somente o neurônio vencedor tem os pesos das suas conexões de entrada reforçadas. A competição entre as células ocorre num plano celular especial, denominado Plano de Seleção de Semente – PSS (*Seed-Selection Plane*, em inglês) (Fukushima, 1982; Cardoso e Wichert, 2010).

O treinamento de uma determinada camada-S, se inicia sem nenhum plano celular já treinado, apenas com o PSS. Ao ser apresentada uma imagem de entrada, os neurônios do PSS, cada um numa posição diferente, competem entre si, e aquele que tiver o maior valor na saída é a célula vencedora, e os pesos do mesmo são reforçados. Cria-se então um novo plano celular, com essa célula vencedora, em todas as posições. Esse procedimento termina quando não for possível obter mais nenhum neurônio vencedor, com as amostras de treinamento consideradas. Quando isso acontece, o treinamento avança para o estágio seguinte, iniciando uma nova etapa de determinação de neurônios vencedores na nova camada-S, usando como entradas as matrizes de características de células-C, resultantes do estágio anterior.

2.2 LeNet

O LeNet é uma rede neural convolucional composta por 7 camadas, possuindo parâmetros ou pesos treináveis, proposta em 1998 por LeCun et al., inspirada no Neocognitron, para o reconhecimento de dígitos manuscritos e impressos (LeCun et al., 1998). Na estrutura original, a camada de entrada recebe imagens com resolução 32x32.

A primeira camada convolucional C1 tem seis planos celulares 28x28, contém 156 parâmetros treináveis, sendo 26 em cada plano, 122,304 conexões, e campos receptivos 5x5. A primeira camada de *pooling* denotado de subamostragem, S2, contém 6 planos, 14x14, 12 parâmetros treináveis, sendo 2 parâmetros em cada plano e 5,880 conexões, considerando 5 conexões por célula, 4 entradas e um viés, dando um total de 980 conexões por plano celular. A segunda camada convolucional C3 contém 16 planos 10x10, 1,516 parâmetros treináveis, que se distribuem em vários planos e 151,600 conexões. A segunda camada de *pooling* S4 contém 16 planos 5x5. Cada unidade em S4 é conectada ao campo receptivo correspondente de tamanho 2x2. Contém 32 parâmetros treináveis, sendo dois em cada plano e 2,000 conexões. A terceira camada convolucional C5 de tamanho 1x1 tem 120 planos e 48,120 parâmetros treináveis, sendo 400 conexões de entrada e um viés, por plano. A camada F6 contém 84 unidades e é totalmente conectada à camada C5, tendo 10.164 parâmetros treináveis, sendo 120 parâmetros de entrada e um de viés, distribuídos em 84 unidades. O treinamento das células no LeNet é realizado usando o algoritmo de retro-propagação (LeCun et al., 1998).

2.3 AlexNet

AlexNet é uma rede neural convolucional, originalmente criada por Alex Krizhevsky, Geoffrey Hinton, e Ilya Sutskever, inspirada na rede LeNet (Krizhevsky et al., 2012).

AlexNet é formada por 8 camadas sendo, cinco camadas convolucionais (Conv 1, Conv 2, Conv 3, Conv 4 e Conv 5) e em cada uma dessas camadas é utilizada a função *ReLU* para ativação dos neurônios. Após uma camada convolucional, a rede contém a camada de *pooling*, com a operação *Maxpooling*. Essa operação de *Maxpooling* é realizada em janelas dentro da matriz de características anterior, sendo que o deslocamento de uma janela para outra (*stride* em inglês), pode variar de uma camada para outra. Além disso, quando a região corresponde a uma borda da matriz, existe um preenchimento (*padding*) das entradas vazias. Após essas camadas, a rede é composta por três camadas totalmente conectadas, sendo na camada de saída a implementação da função *softmax*. A última camada totalmente conectada resulta em 1000 classes. A rede AlexNet, adota a função de ativação ReLU, para realizar o treinamento várias vezes mais rápido que outras funções de ativação como por exemplo tangente hiperbólica, utilizada em LeNet (Krizhevsky et al., 2012).

3 Banco de imagens

As imagens utilizadas nos experimentos preliminares foram extraídas da base de dados da Universidade de Essex (Essex, 2019). Esses dados são mantidos em quatro diretórios (faces94, faces95, faces96, grimace), em ordem crescente de dificuldade. As imagens são armazenadas em RGB de 24 bits, formato JPEG. O número total de indivíduos, entre masculinos e femininos, com e sem artefatos como óculos, barbas e bigodes, é de 395, sendo 20 imagens por indivíduo. A iluminação é artificial, misturando lâmpadas de tungstênio e fluorescentes. A figura 5 ilustra os três tipos de imagens utilizados nos experimentos: (a) imagens da região ocular, (b) imagens de faces completas e (c) imagens da região ocular com acessórios.

(a) Imagens da região ocular	
(b) Imagens de Faces completas	
(c) imagens da região ocular com acessórios	

Figura 5. Amostras de imagens faciais utilizadas nos experimentos.

4 Experimentos

Os experimentos foram realizados de modo a avaliar a operacionalidade para diferentes tipos de amostras. Para a avaliação dos resultados foi considerada a taxa de acerto, conhecida como sensibilidade, que é a relação VP/P, onde VP representa os verdadeiros positivos e P é o total de amostras positivas. Além disso, foi calculado o índice Kappa, proposto por Cohen (Cohen, 1960). A tabela 1 mostra os resultados obtidos.

Nos experimentos usando Neocognitron foram feitos ajustes nos limiares $\theta 1$, $\theta 2$ e $\theta 3$, que determinam o nível de intensidade das entradas para o disparo, nos três estágios da rede, e além disso, foram realizados testes com amostras distintas e não distintas para as fases de treinamento e reconhecimento.

Tabela 1 – Experimentos usando Neocognitron

Exp	1	2	3	4	5	6
Tipo	ocular	completa	acessórios	ocular	completa	acessórios
$\theta 1$	0,97	0,97	0,97	0,90	0,97	0,97
$\theta 2$	0,93	0,97	0,97	0,45	0,97	0,97
$\theta 3$	0,33	0,97	0,97	0,15	0,97	0,97
num.classes	5	5	5	10	7	7
num.amostras	5	5	5	15	7	7
Distintas	Não	Não	Não	Sim	Sim	Sim
VP	24	25	25	15	15	10
FP	1	-	-	30	1	6
Taxa acerto	0,96	1	1	0,3	0,6	0,4
Kappa	0,142	0	0	0,142	0	0

Nos experimentos usando LeNet-5 e AlexNet foram ajustados o número de épocas de treinamento. Além disso, foram realizados testes com amostras distintas e não distintas para as fases de treinamento e reconhecimento. As tabelas 2 e 3 mostram os resultados obtidos para as duas redes, respectivamente.

Tabela 2 – Experimentos usando LeNet-5

Exp	7	8	9	10	11	12
Tipo	ocular	completa	acessórios	ocular	completa	acessórios
Épocas	20	50	50	50	50	50
Num.classes	5	5	5	10	7	7
Num.amostras	5	5	5	15	7	7
Distintas	Não	Não	Não	Sim	Sim	Sim
VP	20	20	19	33	7	7
FP	-	-	1	12	2	2
Taxa acerto	0,8	0,8	0,76	0,66	0,28	0,28
Kappa	0,017	-0,48	-0,48	0,03	0,163	0,266

Tabela 3 – Experimentos usando AlexNet

Exp	13	14	15	16	17	18
Tipo	ocular	completa	acessórios	ocular	completa	acessórios
Épocas	10	10	10	10	10	10
Num.classes	5	5	5	10	7	7
Num.amostras	5	5	5	15	7	7
Distintas	Não	Não	Não	Sim	Sim	Sim
VP	5	19	25	37	8	8
FP	15	5	-	10	1	1
Taxa de acerto	0,2	0,76	1	0,74	0,32	0,32
Kappa	0	0,12	1	0,359	0,11	0,96

5 Análise dos Resultados e Considerações finais

Neste artigo, foram considerados três exemplos de redes para o reconhecimento de faces, por imagens faciais parciais, completas e com acessórios. Para os testes usando amostras distintas da região ocular da face e face com acessórios observou-se que a rede convolucional AlexNet obteve melhor desempenho no índice Kappa. As redes LeNet-5 e AlexNet alcançaram desempenho melhor com amostras não distintas da face completa e face com acessórios respectivamente. Observa-se que o número de amostras utilizadas, tanto para treinamento, como para testes, foi reduzido, resultando-se em taxas de acerto relativamente baixas, principalmente para o Neocognitron. Os coeficientes Kappa obtidos foram também de baixa significância. Para trabalhos futuros, deverá ser considerado o aumento do número de amostras por indivíduos, com outros bancos de imagens faciais, afim de melhorar o desempenho das redes, além de experimentos com diferentes tipos de capturas das imagens de regiões oculares, por exemplo, considerando somente um dos olhos.

Referências

Bengio, Y., and LeCun, Y., 2007 Scaling Learning Algorithms Towards AI, MIT Press

Cardoso, A. e Wichert, A., 2010 Neocognitron and the Map Transformation Cascade, Neural Networks., vol. 23, no. 1, pp. 74–88.

Cohen, J., 1960 A Coefficient of Agreement for Nominal Scales – Educational and Psychological Measurement, Educational and Psychological Measurement , Vol. 20, no. 1, pp.37–46.

Essex, 2019. Essex – Description of the Collection of Facial Images. Disponível em: https://dces.essex.ac.uk/mv/allfaces/. Acessado em janeiro de 2019.

Fukushima, K., 1980 Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biol. Cybernet. 36, pp. 193–202.

Fukushima, K., 1982 Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position, Pattern Recognition, Vol 15, N. 6,

Krizhevsky, A.,Sutskever, I. e Hinton, G. E. 2012, ImageNet Classification with Deep Convolutional Neural Networks

Lecun, Y., Bengio, Y., Bottou, L., e Haffener, P., 1998 Gradient-based Learning Applied to Document Recognition. In Proceedings of the IEEE, pp.2278–2324.

Saito, J. H.; Carvalho, T. V.; Hirakuri, M.; Saunite, A.; Ide, A. N. e Abib, S. 2005 Using CMU-PIE Human Face Database to a Convolutional Neural Network - Neocognitron. - In: Proceedings of European Symposium on Artificial Neural Networks, pp. 491-496.