Incorporando Simetria Rotacional em Redes Neurais Convolucionais Para Reconhecimento de Caracteres

Luiz F. S. Eugênio dos Santos[1]
[1]Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)
São Carlos, São Paulo - Brasil

Resumo

Com o aumento da capacidade computacional aliado à disponibilidade de grandes volumes de dados, uso de redes neurais artificiais vem se tornando cada vez mais popular em diversos domínios. Com o grande número de redes voltadas para tipos diferentes de dados, é conhecida a dificuldade de se estabelecer um formalismo que unifique modelos que operam em estruturas diferentes. Recentemente, o campo de estudo denominado *Deep Learning Geométrico* vem obtendo bons resultados ao explorar a simetria inerente de diferentes tipos de dados na escolha das camadas e funções utilizadas na composição de redes neurais, se destacando em especial no domínio de redes neurais convolucionais (CNNs), redes neurais em grafos (GNNs), bem como outros modelo em conjuntos e até mesmo variedades. Nesse contexto, o presente trabalho busca incorporar camadas convolucionais invariantes à rotação em redes neurais convolucionais clássicas, comparando sua performance em diferentes conjuntos de dados de reconhecimento de caracteres. Tomando como *baseline* os resultados das redes após o treinamento em cada conjunto de dados sem *data augmentation*, é feita a comparação com a rede Spatial Transformer, que implementa transformações que tornam os resultados mais invariantes à rotação dos dados de entrada. A avaliação é feita em diferentes conjuntos de teste, representando ângulos de rotação distintos em relação às imagens originais.

Introdução

Um ponto em comum em diversas arquiteturas de redes neurais profundas mais recentes é seu elevado número de parâmetros a serem treinados, necessitando cada vez mais de grandes conjuntos de dados para seu treinamento bem como de uma elevada capacidade computacional até mesmo para sua operação. Embora a eficácia dessa abordagem até certo ponto possa ser verificada pela crescente performance das redes neurais modernas em diversos domínios, a formalização de uma teoria mais robusta que unifique as arquiteturas que operam em diferentes tipos de dados é interessante tanto para se obter maior interpretabilidade como para que possamos relacionar o conhecimento construído em áreas diferentes.

Nesse contexto, a exploração de características geométricas dos dados vem se mostrando um passo importante na elaboração de novos modelos de redes neurais artificiais para eles. Enquanto alguns exemplo mais notáveis sejam as redes DeepSets (Zaheer et al., 2018), que trata da ausência de uma ordem em conjuntos ao aplicar operações invariantes e equivariantes à permutação (abordagem que também é encontrada em arquiteturas de redes neurais em grafos (Veličković et al., 2018; Wu2, 2021)), outros modelos naturalmente tratam outras características geométricas dos dados sobre os quais operam.

Um exemplo disso são as *Fully Convolutional Networks*, como é o caso da U-Net (Ronneberger et al., 2015), em que a invariância à translação e reflexão é tratada naturalmente ao usarmos os filtros, como ilustrado na Figura 1

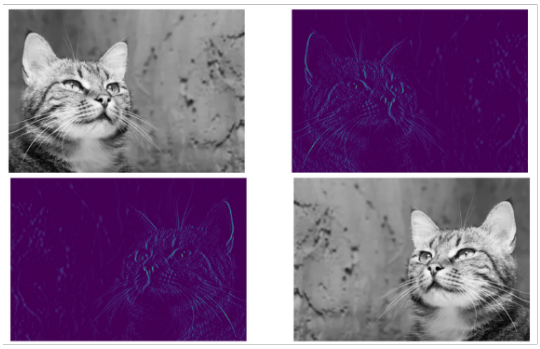

Figura 1: Exemplo de como filtros aplicados em imagens são uma operação invariante à translação e reflexão. Ao aplicarmos uma dessas operações seguida de um filtro, obtemos o mesmo resultado de fazer o inverso, aplicando primeiro o filtro e logo em seguida uma operação de reflexão ou translação.

Uma formalização dessa perspectiva de abordagem na escolha operações que melhor respeitam as simetrias naturais dos dados pode ser obtida através do estudo de grupos e teoria de representações, sendo que sua aplicação direta no cenário de aprendizado profundo vem sendo estudada na recente área de estudo denominada *Geometric Deep Learning* (Bronstein et al., 2017), e já tem obtido bons resultados em uma vasta gama de domínios.

Mais especificamente no contexto de processamento de imagens e redes neurais convolucionais, apesar da já mencionada invariância à translação e relaxação das operações de convolução, a invariância a rotação não ocorre, fazendo inclusive com que a incorporação de imagens rotacionadas seja uma etapa importante no processo de *Data Augmentation*, por exemplo. Para tratar isso, diversas arquitetura derivadas das redes neurais convolucionais clássicas foram propostas, obtendo bons resultados (Mo and Zhao, 2022; Kim et al., 2020; Weiler and Cesa, 2021).

No presente trabalho, usaremos a rede Spatial Transformer (Jaderberg et al., 2015), que além das classes, busca aprender transformações que tornam os resultados mais invariantes à transformações de rotação e translação. Conforme descrito nas secções subsequentes, serão usadas redes convolucionais clássicas como *baseline*, sendo estas treinadas e comparadas ao modelo invariante à transformações para a tarefa de classificação de caracteres em diferentes conjuntos de dados. Todos os códigos desenvolvidos durante o presente trabalho estão estão disponíveis para uso público e contribuições [1].

Metodologia e Avaliação

Após o treinamento das redes com os conjuntos de treino dos *datasets*, será feita a avaliação da acurácia na classificação dos conjuntos de teste após a aplicação de diferentes ângulos de rotação, com variação de 10^o entre 0^o e 350^o. Como demonstrado nos trabalhos citados (Mo and Zhao, 2022; Jaderberg et al., 2015), é esperado considerável ganho de acurácia nas imagens rotacionadas, como mostrado na Figura 2.

No trabalho original, testes como os descritos são feitos nos conjuntos de dados MNIST (Deng, 2012), NWPU VHR-10 (Su et al., 2019, 2020) e MTARSI (Wu et al., 2020), utilizando as redes VGG16 (Simonyan and Zisserman, 2015), ResNet18 (He et al., 2015), DenseNet121 (Huang et al., 2016a) e suas equivalentes RIC-VGG16, RIC-ResNet18 e RIC-DenseNet40, sendo que após a incorporação de imagens rotacionadas, as versões RIC consistentemente apresentaram melhores resultados, como ilustrado na Figura 3

Para os testes do presente trabalho usaremos as versões não invariantes à rotação das redes AlexNet (Krizhevsky et al., 2012), DenseNet121 (Huang et al., 2016b) e ResNet50V2 (He et al., 2016), que serão treinados do zero em cada *dataset* e comparados com a rede Spatial Transformer (Jaderberg et al., 2015). Os conjuntos de dados escolhidos para avaliação são os seguintes: English Handwritten Characters (de Campos et al., 2009a), Kannada Handwritten Characters (de Campos et al., 2009b) e Kuzushiji-MNIST (Clanuwat et al., 2018). Com isso, buscamos estender a avaliação para caracteres de alfabetos diferentes.

[1] https://github.com/LFRusso/SCC0910

Methods	Input Size	Original Test Set	Rotated Test Set
ORN[36]	32×32	99.42%	80.01%
RotEqNet[8]	28×28	99.26%	73.20%
G-CNN[10]	28×28	99.27%	44.81%
H-Net[9]	32×32	99.19%	92.44%
GA-CNN[27]	28×28	95.67%	93.29%
B-CNN[46]	28×28	97.40%	88.29%
E(2)-CNN[47]	29×29	98.14%	94.37%
CNN	32×32	99.55%	45.42%
DEF-CNN[12]	32×32	**99.67%**	46.97%
RIC-CNN	32×32	99.02%	**95.52%**

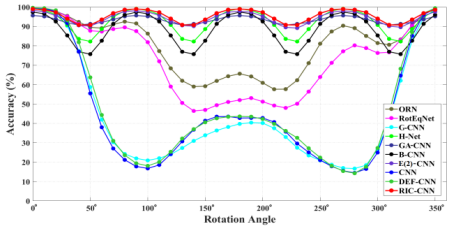

Figura 2: Acurácia no conjunto de dados MNIST de diferentes redes que visam ou não a incorporação de invariância à rotação ou translação às imagens. No gráfico inferior a variação da acurácia em teste para diferentes ângulos de rotação. Fonte: Mo and Zhao (2022)

Training Data	10×100=1K	10×60=0.6K	10×30=0.3K
VGG16	91.64%	87.53%	82.33%
RIC-VGG16	**96.06%**	**93.15%**	**91.65%**
ResNet18	95.84%	93.08%	89.10%
RIC-ResNet18	**98.96%**	**98.32%**	**95.30%**
DenseNet40	97.28%	96.27%	94.30%
RIC-DenseNet40	**99.32%**	**98.89%**	**97.31%**

Figura 3: Avaliação da performance das redes escolhidas na tarefa de classificação para o conjunto de dados NWPU VHR-10, considerando tamanhos de conjunto de treinamento diferentes. Fonte: Mo and Zhao (2022)

Resultados e discussões

Para os testes, todas as redes são construídas de forma à tomarem como entrada imagens em escala de cinza de dimensão 227x227. O treinamento foi feito ao longo de 50 épocas, com *batches* de tamanho 32. Uma vez que o principal objetivo é avaliar a performance das redes selecionadas quanto à sua invariância à rotações, nenhum tipo de pré-processamento dessa natureza foi aplicado durante a etapa de *data augmentation*. Todas as redes foram inicializadas aleatoriamente e treinadas exclusivamente para cada conjunto de dados.

Á princípio, todas as redes foram testadas com os conjuntos de dados originais, sem a aplicação de nenhuma transformação. Os resultados na tarefa de classificação para as diferentes redes podem ser vistos na Figura 4.

Em sequência, todas as imagens do conjunto de teste foram rotacionadas aleatoriamente com um ângulo entre 0^o e

	English	KMNIST	Kannada
AlexNet	0.520	0.953	0.892
DenseNet121	**0.651**	**0.984**	**0.941**
ResNet50V2	0.511	0.977	0.901
STN	0.327	0.929	0.927

Figura 4: Acurácia das diferentes redes nos conjuntos de dados originais, sem o uso de transformações de rotação. Pode-se notar um desempenho pior da rede Spatial Transformer quando comparada com as demais usadas.

360^o. Os testes foram repetidos e os resultados podem ser observados na Figura 5.

	English	KMNIST	Kannada
AlexNet	0.082	0.621	0.552
DenseNet121	0.127	0.588	0.600
ResNet50V2	0.111	0.691	0.618
STN	**0.267**	**0.811**	**0.698**

Figura 5: Acurácia das diferentes redes nos conjuntos de dados originais após a aplicação de rotações aleatórias aos dados. A acurácia da rede Spacial Transformer supera todas as demais nos três conjuntos de teste.

Por fim, todas as figuras do conjunto de teste foram rotacionadas igualmente, com ângulos variando em 10^o, entre 0^o e 350^o. Um gráfico comparando o comportamento de todos os modelos para as diferentes rotações é ilustrado na Figura 6.

Podemos observar que, apesar do desempenho inferior da rede Spatial Transformer para os dados originais, de fato há um considerável ganho de performance nos casos em que há rotação das imagens de entrada. Isso é especialmente interessante no contexto de fotos aéreas, por exemplo, em que os objetos do conjunto de dados podem estar em diferentes orientações, ou em casos em que a aplicação de *data augmentation* é inviável.

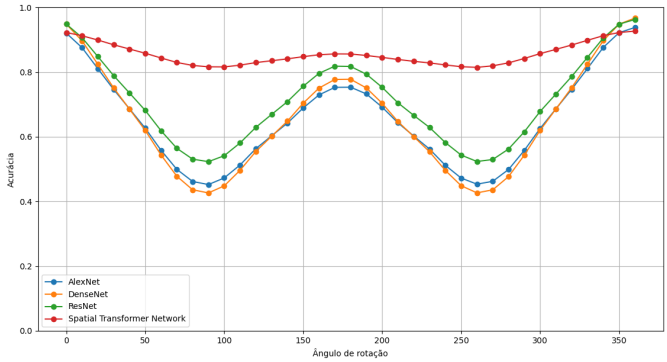

Figura 6: Acurácia das quatro redes no conjunto de dados de teste KMNIST. É possível notar que a performance da rede Spacial Transformer é menos sensível ao ângulo de rotação das imagens, mostrando estar mais próxima à invariância à rotações.

Referências

(2021).

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. (2018). Deep learning for classical japanese literature. cite arxiv:1812.01718Comment: To appear at Neural Information Processing Systems 2018 Workshop on Machine Learning for Creativity and Design.

de Campos, T. E., Babu, B. R., and Varma, M. (2009a). Character recognition in natural images. In *Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal*.

de Campos, T. E., Babu, B. R., and Varma, M. (2009b). Character recognition in natural images. In *Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal*.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. *CoRR*, abs/1603.05027.

Huang, G., Liu, Z., and Weinberger, K. Q. (2016a). Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*.

Huang, G., Liu, Z., and Weinberger, K. Q. (2016b). Densely connected convolutional networks. *CoRR*, abs/1608.06993.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). Spatial transformer networks. *CoRR*, abs/1506.02025.

Kim, J., Jung, W., Kim, H., and Lee, J. (2020). Cycnn: A rotation invariant cnn using polar mapping and cylindrical convolution layers.

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.

Mo, H. and Zhao, G. (2022). Ric-cnn: Rotation-invariant coordinate convolutional neural network.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Su, H., Wei, S., Liu, S., Liang, J., Wang, C., Shi, J., and Zhang, X. (2020). Hq-isnet: High-quality instance segmentation for remote sensing imagery. *Remote Sensing*, 12(6).

Su, H., Wei, S., Yan, M., Wang, C., Shi, J., and Zhang, X. (2019). Object detection and instance segmentation in remote sensing imagery based on precise mask r-cnn. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1454–1457.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks.

Weiler, M. and Cesa, G. (2021). General $e(2)$-equivariant steerable cnns.

Wu, Z.-Z., Wan, S.-H., Wang, X.-F., Tan, M., Zou, L., Li, X.-L., and Chen, Y. (2020). A benchmark data set for aircraft type recognition from remote sensing images. *Applied Soft Computing*, 89:106132.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., and Smola, A. (2018). Deep sets.