Uso da técnica de Single Shot Detection para melhorar a precisão da predição de Convolution Neural Network

Felipe Ferreira
a15023@bcc.unifal-mg.edu.br

Humberto Brandão
humberto.brandao@gmail.com

Dezembro, 2019

Resumo

Com o avanço da tecnologia nos tempos atuais e a medida que as técnicas de aprendizado de máquina se tornam cada vez mais conhecidas, tem se tornado muito comum o uso de placas de processamento gráfico para se desenvolver usando bases de dados em que as informações estão em formato de imagem, com isso, é cada vez mais importante encontrar métodos que melhorem a precisão das predições, um dos objetivos principais quando se fala em aprendizado de máquina. Single Shot Detection é uma técnica que permite a identificação de diferentes objetos em uma mesma imagem, tornando possível aumentar a quantidade de dados que podem ser utilizados para treinar uma rede de convolução, com base nas semelhanças encontradas nas imagens e gerando novos grupos de aprendizado. Isso permite que as redes sejam treinadas com as características mais importantes, como poderemos ver nos resultados, um ganho de precisão de por volta de 4% no desempenho final dos modelos.

Palavras-chave: Machine Learning, Deep Learning, Residual Network, Convolution Neural Network, Single Shot Detection, Data Augmentation

1 Introducão

Ao trabalhar com base de dados, recorrentemente nos deparamos com os mais diversos tamanhos, dependendo do tipo de problema com o qual estamos lidando. Em bases de imagens, isto não é diferente. Na busca por métodos de ampliar a confiabilidade dos resultados, até mesmo nas menores bases, foi criado o método de 'Data Augmentation' que visa aumentar artificialmente um conjunto de dados, baseando-se nos dados já existentes [6, 13]. Na tentativa de utilizar este método de uma maneira mais precisa e menos aleatória, a técnica de Single Shot Detection (SSD) destaca-se por detectar objetos contidos em uma imagem indicando sua exata posição. Deste modo, utilizando um algoritmo, é possível cruzar objetos que apareçam em múltiplas imagens da base de dados, o que permite realizar assim treinamentos mais específicos de Convolutional Neural Networks (CNN). Com múltiplas CNN treinadas, o passo seguinte é utilizar de métodos como o 'Ensemble' para mesclar os diversos resultados obtidos em um único resultado mais preciso que qualquer outro modelo simples.

2 Trabalhos relacionados

Na busca por melhores resultado em uma predição, muitas técnicas podem ser utilizadas para o ganho de performance, assim como Krizhevsky [6] utiliza de 'Data Augmentation' em seu trabalho. Ainda é possível citar os métodos de 'Batch Normalization' [4] e 'Dropout' [3]. Assim como Viola e Jonas [11], que utilizam de um detector de objetos para focar em identificar faces e obter resultados muito mais rápidos, identifica-se possível utilizar de algoritmos genéricos para se realizar experimentos mais específicos. Os experimentos realizados são baseados em estudo de paletas de cores como descrito por Kuleshova [7] e sua comparação as característica de rosto.

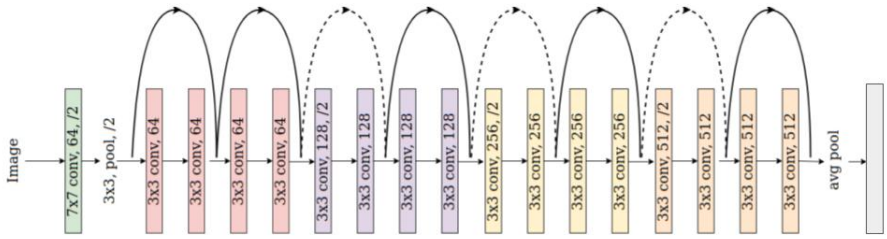

Figura 1: Exemplo do modelo ResNet de 18 camadas

3 Métodos

Ao se iniciar uma análise de base de dados, é necessário definir alguns fatores importantes como o padrão de comparação dos resultados obtidos. Primeiramente, é necessário realizar a definição de uma métrica de validação estabelecendo critérios de pontuação para cada acerto e erro que obtivermos no treinamento do modelo. Neste caso, os resultados obtidos neste trabalho foram baseados na métrica de precisão, que leva em consideração a razão entre os acertos e o total de amostras testadas.

Além da métrica é importante definir o modelo a ser utilizado durante todo o projeto, de modo a padronizar como o aprendizado será realizado. Tratando-se de base de imagens, temos como opção utilizar uma CNN pré treinada que possui diversos modelos, para treinamento com diferentes quantidades de camadas de profundidade, chamada de Residual Network (ResNet) [2]. É interessante verificar se existe melhoria nos modelos para as diferentes quantidades de camadas, pois em algumas delas poderia haver um aprendizado mais aprofundado quando trabalhado na imagem original, que quando testado nas separadas partes da imagem, não haveria mudança. Neste trabalho, foram feitas analises nos modelos de 18, 34 e 50 camadas, pois são os modelos ResNet mais simples para serem usados como base dos resultados, na Figura 1 é possível ver como funciona cada camada da ResNet18. Para preparar esses modelos, foram realizados 2 treinamentos, o primeira para imagens no tamanho 128x128 e em seguida feito o redimensionamento das mesmas para 256x256, esses tamanhos foram escolhidos por meio de tempo de execução e testes anteriores, os parâmetros possuem os seguintes padrões:

- Modificação nas camadas bloqueado, 25 iterações de treinamento e taxa de aprendizado de 0.01
- Modificação nas camadas liberado, 25 iterações de treinamento e taxa de aprendizado de 0.0001

O bloqueio das camadas torna o modelo mais linear exigindo que passe por todas as camadas, a liberação o torna mais flexível permitindo pular dependendo dos resultados que seriam obtidos na camada. A taxa de aprendizado quanto mais baixa, mais lento o algoritmo aprende, porém mais confiável se torna esse aprendizado.

Definidas as métricas e o modelo, é possível realizar uma primeira análise da base de dados partindo-se do princípio que a mesma não tenha sofrido nenhuma modificação até então, para que possamos definir uma baseline para o problema, que virá a ser usada como ponto inicial da comparação das melhorias que venham a ocorrer durante a execução do projeto. Cada vez que essa análise é realizada utilizando a CNN, obtemos um resultado muito próximo, porém diferente do anterior. Para resolver esse problema e diminuir a chance de obter um resultado

Figura 2: Marcações geradas pelo algoritmo de SSD em algumas imagens da base de dados utilizada

com grande variância, o mesmo teste é realizado diversas vezes, para este caso é decidido 5 como uma quantidade plausível por questão de tempo de execução pois a melhoria que seria obtida com valores maiores seria muito baixa comparada ao tempo de execução necessário, em seguida os resultados são agrupados, realizando a média dos mesmos.

Para garantir a precisão do modelo, é preciso utilizar um método de validação cruzada, neste caso, como a quantidade de imagens é muito pequena comparada à quantidade de características que estão sendo utilizadas, a opção que se torna mais confiável, é a técnica de 'Leave One Out' que consiste na remoção de apenas uma imagem da base de treinamento, a qual é utilizada ao fim do processo como um teste de confiabilidade. Isto é feito uma vez para cada imagem, até que todas sejam utilizadas.

Para dar maior diversidade à base de dados, é interessante aumentar a quantidade de imagens existente nela, mas em muitos contextos isso é de difícil realização. Por isso, o algoritmo de SSD se torna uma alternativa interessante, uma vez que identifica diferentes objetos dentro de uma mesma imagem permitindo a criação de novas imagens a partir dos objetos detectados. O algoritmo de SSD possui como saída um código que contendo o objeto identificado, 2 valores para x e 2 valores para y, que representam o inicio e o fim da imagem em pixels, e um valor que representa a porcentagem de certeza que o algoritmo possui de que o objeto é realmente o que foi identificado. Com estes dados, é possível, baseado nos objetos mais encontrados, realizar os cortes nas imagens usando as posições x e y em que se encontram e assim gerar novas imagens para uma base de dados diferente da existente.

Com novas imagens faz-se necessário a reformulação da base de treinamento. Para isso, é importante categorizar os objetos identificados e agrupar todos os que fazem parte da mesmo categoria podendo gerar até mais de uma base de dados nova. A partir disso, novos treinamentos devem ser aplicados, seguindo as métricas e modelos definidos previamente.

Em busca de um único resultado sobre cada uma das imagens da base, é aplicado a técnica de 'Ensemble' para unificação dos múltiplos resultados. Esta técnica se baseia na média dos

Figura 3: Alguns exemplos do resultado da criação de novas bases de dados

	Autumn	Spring	Summer	Winter	Alvo	Predição
Imagem 1	3,2058 %	28,6233 %	1,0679 %	67,8579 %	Winter	Winter
Imagem 2	93,9559 %	3,1236 %	1,6618 %	2,3361 %	Winter	Autumn

Tabela 1: Resultado das 2 primeiras imagens do modelo gerado pela CNN ResNet-18

	ResNet-18	ResNet-34	ResNet-50
Taxa de Acerto	65,3846 %	67,3077 %	71,1538 %

Tabela 2: Taxa de acerto das predições das imagens originais em cada modelo utilizado

resultados obtidos por cada um dos treinamentos anteriores, com pesos igualmente distribuídos.

4 Resultados

Os resultados obtidos nos experimentos foram baseados em uma base de dados de 104 rostos femininos distintos, que estão categorizados de 4 maneiras diferentes (winter, autumn, summer, spring) que fazem referência à paleta de cores que mais se adéqua ao estilo de rosto da pessoa. Cada uma das linhas, que se refere a uma imagem, gera como saída a porcentagem para cada uma das categorias, sendo essa porcentagem independente uma da outra. O alvo é a qual categoria a imagem realmente pertence e a predição é o resultado gerado pelo modelo como no exemplo da Tabela 1

Com os dados da Tabela 2 sendo utilizados como ponto de referência do estudo, é possível então realizar uma comparação com os resultados obtidos após a separação das categorias e re-processamento dos modelos com os parâmetros pré estabelecidos.

Os dados da Tabela 3 demonstram os resultados quando trabalhados de forma separada, realizando um experimento para cada categoria de objeto reconhecido. É possível notar que os resultados, quando é feita a análise apenas dos olhos, mostram ser esta a característica que mais possui significância para este problema. Segundo Mary Spillane (1995, p. 29, tradução nossa)[1] "não é uma característica importante -a cor do seu cabelo, olhos ou tom de pele, mas a imagem que as três criam juntas que nos fornece um guia para o nosso melhor visual sazonal", por isso os corte isolados não fornecem tanta precisão quanto utilizando a imagem original.

Na Tabela 4 temos a representação da predição final para cada modelo, em que é incorporado no resultado as predições de cada característica, utilizando o método de 'Ensemble'. O uso deste método permitiu, então, gerar resultados melhores quando selecionada a categoria que possui maior porcentagem de certeza dentre todas as classes.

[1] "it's not one feature -the color of your hair, eyes, or skintone that's important, but the picture the three create together that provides us whit a guide to our best seasonal look"

	ResNet-18	ResNet-34	ResNet-50
Olhos	53,8462 %	55,7692 %	58,6538 %
Boca	51,9231 %	54,8077 %	55,7692 %
Nariz	49,0385 %	50,9615 %	51,9231 %

Tabela 3: Taxa de acerto das predições de cada categoria de imagem para cada modelo utilizado

	ResNet-18	ResNet-34	ResNet-50
Taxa de Acerto	68,2692 %	71,1538 %	75,9615 %

Tabela 4: Taxa de acerto das predições após mesclagem dos resultados das categorias

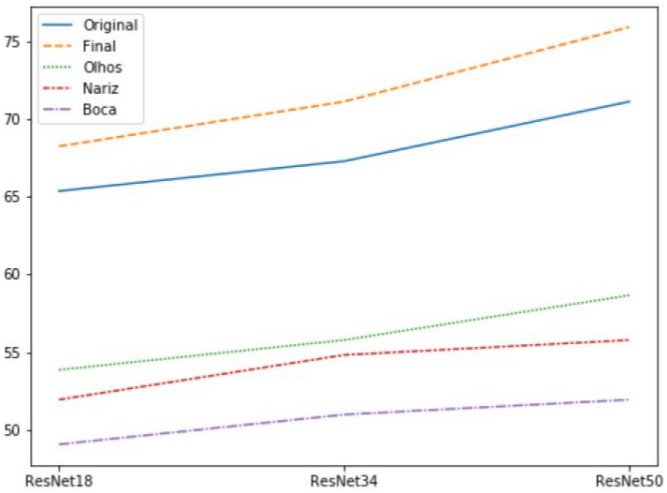

Figura 4: Gráfico dos resultados obtidos de cada modelo em cada uma das bases de dados.

5 Discussão

A melhoria nos resultados das predições é evidente ao se realizar a categorização das partes e posterior junção de resultados, em contraste ao modelo com imagens sem modificação.

Desde o princípio a maior preocupação foi garantir a confiabilidade nos resultados obtidos ao se trabalhar com bases de pequenas dimensões. Uma maneira de assegurar isto, foi utilizando o método de validação cruzada "Leave One Out". Segundo Ron Kohavi (1995, p. 6, tradução nossa)[2] "validação cruzada k-fold com valor moderado de k reduz a variância enquanto aumenta o viés". Kohavi ainda complementa a cerca da base com poucos dados:[3] "À medida que k diminui e o tamanho da amostra diminui, há variação devido à instabilidade do treinamento". O fato de realizar múltiplas execuções do mesmo modelo com diferentes sementes e mesclá-los no fim garante ainda maior estabilidade nos testes realizados.

O modelo de CNN escolhido para este problema é devido à sua maior velocidade de conseguir atingir o limite de melhoria da base. Segundo Kaiming He (2016, p. 6, tradução nossa)[4] "a ResNet facilita a otimização, proporcionando uma convergência mais rápida no estágio inicial".

É de se esperar que os resultados do modelo gerado, utilizando as imagens originais apresentados na Tabela 1, seriam melhores que cada uma das características separadas referentes ao mesmo modelo na Tabela 2 devido à quantidade de informações que foi retirada da imagem, com função importante no aprendizado do modelo.Uma vez que essas informações foram separadas e divididas em vários aprendizados específicos distintos, tornaram-se as características que diferenciam realmente as imagens mais perceptíveis no momento do treinamento do modelo, fazendo com que no fim, ao serem mescladas, modelos que possuem maior confiança na predição se destaquem de outros que estão em dúvida de duas ou mais categorias, tornando o resultado mais preciso, assim como mostrado na Tabela 3. É importante também citar que essa melhoria foi observada em todos os modelos em que os testes foram realizados, mostrando assim que os resultados são consistentes.

A Figura 3 permite melhor visualização dos diversos modelos quando comparados entre si, e a melhoria do ensemble dos resultados em relação ao original.

[2] "k-fold cross-validation with moderate k values reduces the variance while increasing the bias"
[3] "As k decreases and the sample sizes get smaller, there is variance due to the instability of the training"
[4] "the ResNet eases the optimization by providing faster convergence at the early stage"

Referências

[1] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012.

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[5] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Svetlana Kuleshova, Alla Slavinska, Oksana Zakharkevich, and Galina Shvets. Image clothing as a perceptual component of clothing design. 2017.

[8] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016.

[9] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[10] Mary Spillane and Christine Sherlock. *Color Me Beautiful's Looking Your Best: Color, Makeup and Style*. Madison Books, 1995.

[11] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[12] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[13] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.