Identificação de Baleias Jubarte pela Cauda

Henrique da Fonseca Simões[1] and João Meidanis[1]

[1] *Instituto de Computação (IC) - Universidade Estadual de Campinas (UNICAMP)*

Baleias jubarte (*Megaptera novaeangliae*) podem ser identificadas pela suas caudas, por meio do formato e padrões de manchas, linhas e listras. Isto possibilita o rastreamento dos indivíduos ao longo do tempo e o entendimento da dinâmica populacional da espécie. Por décadas, esta análise foi feita manualmente por pesquisadores, mas entre novembro de 2018 e fevereiro de 2019, a plataforma *Kaggle* sediou uma competição para a realização automatizada dessa tarefa. Em nosso projeto, realizamos a reprodução dos treinamentos das três soluções melhor colocadas na competição, de forma a verificar a reprodutibilidade dos resultados. Descobrimos que a segunda e terceira soluções são mais reprodutíveis que a primeira, atingindo uma pontuação de 0.97 na métrica MAP@5. Além disso, havia incorreções nos rótulos dos conjunto de dados e também exemplos que dificultavam a identificação dos padrões na cauda. Corrigimos estes problemas, criando um novo conjunto de dados, e obtivemos uma melhora de 1% na solução do segundo e terceiro colocados, no novo conjunto.

I. INTRODUÇÃO

Baleias jubarte (*Megaptera novaeangliae*) foram uma das espécies predominantemente capturadas pela indústria da pesca de baleias entre a década de 1860 e o fim da década de 1900 [13]. Apesar de haver a recuperação da população de baleias após banimento da pesca e a espécie ser declarada ameaçada de extinção, os dados para documentação da recuperação populacional ainda são limitados para algumas subpopulações [15]. No acompanhamento da dinâmica populacional, a identificação dos indivíduos é realizada por meio da análise de fotografias de suas caudas, as quais possuem formato e padrões de manchas, linhas e listras capazes de distingui-los [10]. Entretanto, essa análise foi feita por décadas manualmente por pesquisadores.

De novembro de 2018 a fevereiro de 2019, a plataforma *Kaggle* sediou uma competição para a identificação automatizada de baleias jubarte pela cauda [9], onde os participantes recebiam fotografias das baleias e tinham de determinar as suas classes (identidades). Como ponto de partida de nossas atividades, analisamos as três soluções melhores colocadas, com o objetivo de reproduzir os resultados obtidos pelos competidores. Na sequência, identificamos e corrigimos alguns problemas no conjunto de dados e reavaliamos as soluções. Por fim, testamos duas adaptações na solução de melhor resultado.

II. SOLUÇÕES

Em nossa análise, consideramos as três soluções de melhor classificação no placar privado da competição, que utilizou a métrica MAP@5 para avaliação das predições.

A. Primeira colocação

Criada por Jian Qiao, Peiyuan Liao, Thomas Tilli e Yiheng Wang, esta solução atingiu a primeira colocação na competição, com pontuação 0.97309. A estratégia utilizada consiste em classificar as imagens por meio de uma rede neural profunda, a SENet-154 [6], utilizando extração local e global de características [14]. Outras técnicas também são utilizadas, como aumento de dados por meio de diversas transformações, criação de máscaras e *bounding boxes* para as caudas, aplicação da função de custo *Triplet Loss* [5] — que direciona as representações de exemplos semelhantes a estarem mais próximas entre si do que as de outras classes — e treinamento *a posteriori* de exemplos únicos (*one-shot*) [3].

B. Segunda colocação

Esta solução, projetada por Tao Shen, também utilizou redes neurais profundas para classificação das imagens e atingiu pontuação 0.97208. Como preprocessamento, é realizado corte na imagem para focar somente na cauda, padronização das dimensões das imagens e aumento virtual dos dados. Uma rede suporte então gera um vetor de características, que é mapeado nas classes por meio de uma camada densa. Nas 100 épocas de treinamento da rede, quatro funções de custo são combinadas: *Triplet Loss* [5], *ArcFace Loss* [1], *Cos-*

Face Loss [16, 17] e uma adaptação da *Focal-Loss* [11].

No processo de inferência, os resultados de 10 modelos são combinados para a geração da classificação final. Estes modelos variam na rede suporte — ResNet-101 [4], SE-ResNet-101 [4, 6] ou SE-ResNeXt-101 [6, 18] —, uso de pseudorrótulos e tamanho das imagens (512×256 ou 512×512).

C. Terceira colocação

Jinmo Park projetou uma arquitetura que utiliza aprendizado profundo, atingindo 0.97113 de pontuação e conquistando a terceira colocação na competição. Sua solução também usa aumento virtual dos dados por transformações, *bounding boxes*, além de um procedimento de alinhamento da cauda na imagem. Como rede suporte, é utilizada a DenseNet-121 [7]. Os mapas de características criados por ela são mapeados por meio de uma camada densa para *embeddings* com 512 dimensões, que sumarizam as características da cauda da baleia. Visando um aumento do poder discriminativo da rede, a estratégia *ArcFace* [1] é utilizada no treinamento (500 épocas).

Na inferência, um *embedding* representante de cada classe é criado combinando os *embeddings* de suas imagens do conjunto de treino. Eles são então comparados por meio de similaridade de cosseno ao da imagem a ser classificada. Na submissão final, Park utilizou a combinação de três modelos utilizando essa arquitetura com pequenas variações nos hiperparâmetros e agrupamento de algumas classes de treino.

III. CORREÇÃO DOS DADOS

Com auxílio de *posts* no fórum da competição, notamos incorreções nos rótulos no conjunto de dados, além de exemplos significativamente fora do padrão. Realizamos então uma inspeção dos dados de treino (25361 imagens) para correção dos problemas identificados e criamos testes utilizando os dados corrigidos. Com isso, pudemos reavaliar as soluções e identificar o impacto das mudanças.

A. Rótulos

Encontramos principalmente dois problemas: baleias iguais com rótulos diferentes e baleias identificáveis com rótulo *nova baleia*.

(a) Ângulo da fotografia fora do padrão

(b) Visão da cauda fora do padrão

(c) Parte da cauda submersa

(d) Oclusão causada por aspersão d'água

Figura 1: Exemplos de algumas situações consideradas difíceis.

No primeiro quesito, encontramos 121 pares e 4 trios de classes suspeitas. Confirmamos que 108 pares eram, de fato, duplicatas e todos os trios também o eram. Com isso, escolhemos um dos rótulos para representar a baleia no conjunto de dados corrigido, o que impactou cerca de 700 imagens. Também corrigimos rótulos de imagens isoladas quando necessário.

No segundo quesito, encontramos 336 imagens originalmente na classe *nova baleia* que potencialmente não seriam novas. Dessas, confirmamos que 164 pertenciam a alguma baleia identificável. A correção consistiu, então, na mudança para o rótulo da baleia em questão.

B. Exemplos difíceis

Outra característica identificada no conjunto de dados foi a presença de exemplos com potencial de dificultar substancialmente a predição; por exemplo, caudas que estão apenas parcialmente visíveis na imagem e, em sua maioria, "rotacionadas" em relação ao nível da água. Note que essa dificuldade se deve pela necessidade haver uma generalização não somente em relação à disposição das marcas, mas também à localização e angulação delas na imagem. Assim, decidimos separar exemplos com estas características em um conjunto de teste que denominamos de *difícil*. A Figura 1 ilustra situações encontradas.

Encontramos também situações onde os exemplos não estavam adequados. A situação mais frequente foi a existência de múltiplas caudas em uma mesma imagem. No total, 66 imagens foram identificadas e desconsideradas do conjunto corrigido.

IV. REAVALIAÇÃO DAS SOLUÇÕES

Para a reavaliação com os dados corrigidos, criamos dois conjuntos utilizando uma amostragem aleatória de 20% dos exemplos não-difíceis para teste e os 80% restantes para treino. Um conjunto adicional de teste foi criado com os exemplos difíceis. A Tabela I mostra algumas informações acerca de cada conjunto criado. Um total de 4887 classes diferentes de nova baleia estavam presentes na totalidade dos dados utilizados, sendo que 4353 dessas tinham ao menos um exemplar no conjunto não-difícil. Como algumas classes não tinham exemplares no conjunto de treino mas estavam presentes no conjunto de teste, consideramos a classe correta delas como *nova baleia* na avaliação da predições.

Na reavaliação das soluções, não realizamos *ensembles*. Assim, escolhemos o modelo da segunda solução que obteve a maior pontuação na primeira parte deste estudo, dentre os que utilizam tamanho de imagens 512×256 e sem pseudorrótulos: o ResNet-101 [4]. Para a terceira solução, utilizamos a arquitetura descrita, sem agrupamento de classes e com a redução do número de épocas treinadas para 300, adequando proporcionalmente os marcos do planejador de taxa de aprendizagem.

V. ADAPTAÇÕES

O algoritmo do segundo colocado mostrou algumas vantagens para ser utilizado em detrimento dos demais: (a) atingiu os melhores resultados nos testes que fizemos na primeira etapa do projeto, (b) teve resultados mais estáveis (em relação à primeira solução) e (c) exibiu tempos de treinamento e inferência menores.

Desta forma, utilizamo-lo como base para testar duas adaptações[1]: utilização de normalização de contraste e aplicação da estratégia de Média de Pesos Estocásticos (em inglês *Stochastic Weight Averaging* - SWA) [8].

A. Normalização de Contraste

Contraste — desvio padrão dos pixeis — é um tipo de variação existente nas imagens que pode ser removido, auxiliando a rede neural a não considerá-lo na realização da tarefa em questão [2]. A estratégia já utilizada na segunda solução relacionada com contraste é a subtração do valor médio e divisão pelo desvio padrão por canais de cores com base em estatísticas das imagens da base de dados *ImageNet* [12].

Assim, testamos outras três formas de lidar com contraste: sem alteração do contraste (sem CN), utilizando Normalização Global de Contraste (GCN) e Normalização Local de Contraste (LCN) [2]. Para GCN, utilizamos a formulação original, sem termo regulador de contraste. Para LCN, consideramos o termo regulador $\lambda = 1$, uma janela de 9×9 pixeis e *padding* para manutenção das dimensões da imagem.

B. Média de Pesos Estocásticos

Esta técnica consiste em realizar uma combinação de várias instâncias de um modelo durante seu processo de convergência. Esta combinação é feita por meio da média móvel dos parâmetros. Izmailov e colegas mostram uma diminuição no erro de validação e uma melhora na generalização de diferentes redes neurais, inclusive residuais, com essa técnica, ao utilizar otimização pelo método do Gradiente Descendente Estocástico com uma alta taxa de aprendizagem [8].

Aplicamos essa técnica da 75º época até 100º época, com uma taxa de aprendizagem de 10^{-4} e com *cosine annealing* de 10 épocas.

VI. RESULTADOS

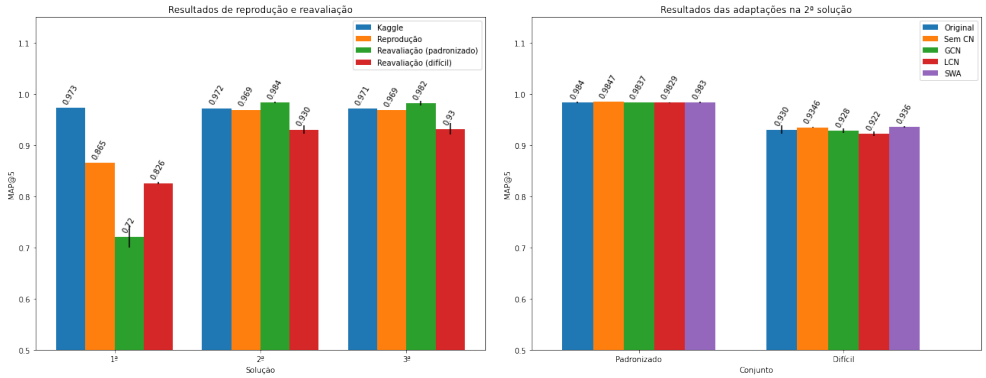
Realizamos os treinamentos de todos os modelos seguindo as orientações fornecidas pelos autores para a reprodução dos resultados. Nas reavaliações e adaptações, cada modelo foi treinado com os dados de treino de cada conjunto e avaliada no dados de teste do conjunto e no conjunto difícil. A Figura 2 mostra os resultados obtidos.

Como pode ser observado, a primeira solução teve um resultado mais precário em todos os nossos experimentos. Percebemos que escolha de pontos de mudança nas fases de treinamento impacta significativamente no resultado do modelo treinado. Isso traz indícios de que seja possível obter resultados melhores do que os nossos com essa arquitetura, se houver mais domínio da seleção desses pontos, aliado a mais poder computacional. Porém, independentemente disso, a solução não se mostra facilmente reprodutível. Percebe-se também que ela foi a única que teve uma pontuação maior no con-

[1] Código disponível em `https://github.com/henriquesimoes/humpback`

	Conjunto 1		Conjunto 2		Conjunto difícil
	Treino	Teste	Treino	Teste	Teste
Classes	3965	1747	3927	1805	1441
Imagens de baleias identificáveis	11161	2792	11139	2814	1869
Imagens de novas baleias	4813	1201	4835	1179	3459
Total de imagens	15974	3993	15974	3993	5328

Tabela I: Número de classes, imagens de baleias identificáveis e de novas baleias em cada conjunto criado com dados corrigidos. O número de classes se refere às classes diferentes de nova baleia que tinham ao menos um exemplo no conjunto em questão.



(a) Reprodução e reavaliação

(b) Adaptações

Figura 2: Resultados dos treinamentos dos modelos (a) nas reproduções das top-3 soluções, avaliadas na plataforma Kaggle, e nas reavaliações no conjunto de dados corrigidos e (b) com adaptações realizadas na solução de segunda colocação, avaliadas no conjunto corrigido.

junto difícil, o que não era esperado. Analisando as predições, percebemos que tal aspecto se deveu ao maior número de novas baleias presentes no conjunto (Tabela I). Isso porque havia uma propensão maior da rede a indicar baleias como sendo novas; o que traz mais acertos no conjunto difícil do que no padronizado.

Apesar disso, a correção dos rótulos e separação dos exemplos difíceis trouxe um efeito positivo na aprendizagem do modelo da segunda e terceira soluções. Em nossos testes, houve uma melhor generalização nos dados padronizados, em relação a submissão dos autores no *Kaggle*, mesmo sem utilizar *ensembles*, com uma diferença absoluta de 0.012 ± 0.001 e 0.011 ± 0.005 pontos para a segunda e terceira soluções, respectivamente. Isso indica que o limitante de melhora da solução projetada estava relacionado com os dados em si e, provavelmente, não tanto com as arquiteturas criadas.

Outro indicativo nesse sentido são os resultados com adaptações na segunda solução. Observa-se que não houve significativo impacto na generalização dos modelos, tanto com alterações no contraste quanto com a Média de Pesos Estocásticos.

VII. CONCLUSÃO

Analisamos as três soluções melhor colocadas na competição *Kaggle* de reconhecimento de baleias jubarte pelas suas caudas. Identificamos que a segunda e terceira soluções são mais facilmente reprodutíveis.

Além disso, percebemos que um dos impeditivos de melhora dos resultados das soluções era o próprio conjunto de dados, em que erros nos rótulos estavam presentes, além exemplos que traziam desafios que iam além do simples reconhecimento do formato e dos padrões de marcas na cauda. Com dados mais bem curados, melhorias nos resultados, como no modelo de segunda solução, parecem ser possíveis, embora modestas.

Experimentamos também adicionar outras duas técnicas na solução mais promissora (a segunda colocada na competição), porém sem resultados expressivos de melhoria.

AGRADECIMENTOS

Este projeto foi apoiado[2] pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo no. 2019/11386-3. Agradecemos ao Instituto de Computação e ao Serviço de Apoio ao Estudante (SAE) pelo apoio institucional dado durante a execução do projeto, assim como o suporte da FAPESP na aquisição sob processo no. 2018/00031-7 da máquina utilizada em nossos experimentos.

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[3] Yandong Guo and Lei Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[8] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[9] Kaggle Team. Humpback Whale Identification: Can you identify a whale by its tail? `https://www.kaggle.com/c/humpback-whale-identification`, 2019. Acessado em: 2020-09-28.

[10] Steven Katona, Ben Baxter, Oliver Brazier, Scott Kraus, Judy Perkins, and Hal Whitehead. Identification of humpback whales by fluke photographs. In *Behavior of marine animals*, pages 33–44. Springer, 1979.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2988, 2017.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal on Computer Vision*, 115(3): 211–252, 2015.

[13] Peter T Stevick, Judith Allen, Phillip J Clapham, Nancy Friday, Steven K Katona, Finn Larsen, Jon Lien, David K Mattila, Per J Palsbøll, Jóhann Sigurjónsson, Tim D. Smith, Nils Øien, and Philip S. Hammond. North Atlantic humpback whale abundance and rate of increase four decades after protection from whaling. *Marine Ecology Progress Series*, 258: 263–273, 2003.

[14] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[15] Peter O Thomas, Randall R Reeves, and Robert L Brownell Jr. Status of the world's baleen whales. *Marine Mammal Science*, 32(2): 682–734, 2016.

[16] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive Margin Softmax for Face Verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[17] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

[18] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017.

[2] As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão da FAPESP.