Análise comparativa de redes para segmentação semântica

Pedro Lima
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
pbsl@cin.ufpe.br

Gabriel Vasconcelos
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
ghv@cin.ufpe.br

***Resumo*—Este documento propõe uma análise comparativa da aplicação de três redes convolucionais para a segmentação semântica, aplicadas ao dataset de ultrasonografias de câncer de mama. A proposta é aprofundar o entendimento das arquiteturas U-net, Residual U-net e FastFCN, em relação ao funcionamento geral delas e também nos resultados que obtivermos da aplicação em um problema real.**

***Palavras-chave*—segmentação, imagem, convolucional**

I. Introdução

O câncer de mama é o tipo mais comum de câncer entre as mulheres, sendo o tipo de câncer com mais diagnósticos, sendo estimados mais de 2,5 milhões diagnósticos em mulheres, além responsável pela morte de aproximadamente 685 mil mulheres, tudo isso apenas no ano de 2020 [1]. Com o grande e rapidamente crescente número de diagnósticos e fatalidades, a necessidade por diagnósticos rápidos, eficientes e confiáveis também vêm aumentado; e uma possibilidade de tipos de diagnósticos que se encaixam nesses critérios são diagnósticos por meio de redes neurais artificiais. Com o campo da visão computacional apresentando cada vez mais modelos, se torna difícil acompanhar quais são os métodos mais precisos, ou com maior custo-benefício, para a detecção, classificação e segmentação de agentes patogênicos no corpo humano. Por meio da análise apresentada neste relatório, serão apresentadas as comparações entre os resultados da segmentação levando em consideração o custo computacional.

II. Objetivo

O objetivo dessa análise é realizar uma comparação entre três modelos de aprendizagem profunda, voltadas para problemas de segmentação diferentes, a U-Net, a Residual U-Net e a FastFCN. A comparação com base em suas previsões e custo computacional será feita para analisar qual a viabilidade de cada uma em um cenário de diagnóstico médico.

III. Justificativa

A motivação para o projeto foi inspirada pelo crescente número de pessoas diagnosticadas com câncer de mama, e consequentemente com a crescente demanda por diagnósticos mais rápidos, baratos, confiáveis e efetivos; que tem como uma opção o uso das redes neurais convolucionais. Com isso, também é fato de que o número de arquiteturas de redes neurais tem crescido; o que implica a necessidade à busca por quais arquiteturas são eficientes, na segmentação, e classificação de células cancerígenas, assim como em seu treinamento e em seu gasto de recursos.

IV. Metodologia

A metodologia e o desenvolvimento do projeto pode ser dividido em 4 etapas: Análise e estudo do banco de dados, a implementação dos modelos CNN, o treinamento e, por fim, testes e análises de resultados.

A. Análise e estudo do banco de dados

Inicialmente, o dataset escolhido para ser o alvo da análise foi um dataset, encontrado no Kaggle, da BACH Challenge 2018 [2], que era dividido em imagens para problemas de classificação e em imagens para segmentação, que seria a parte usada, de acordo com nossa proposta. As imagens para segmentação eram imagens de células mamárias que podiam ser classificadas dentre 4 classes diferentes. Porém a utilização do dataset se provou bastante difícil para o grupo, pois a quantidade total de imagens disponíveis, para treinamento e validação de modelos de segmentação, eram de apenas 30 imagens, todas no formato SVS, que é uma formatação específica para imagens médicas [3]. O grupo se deparou com bastante dificuldade em utilizar as imagens, não tendo sucesso em abrí-las ou em transformá-las para arquivos de imagens mais acessíveis para sua manipulação, como PNG ou JPG, devido à falta de compatibilidade de bibliotecas com o Google Colab, que eram capazes de lidar com arquivos SVS. Além disso, devido ao número limitado de imagens disponíveis no dataset, o grupo questionou o fato de que os modelos poderiam ser treinados bem o suficiente para que comparações relevantes poderiam ser feitas entre eles. Por esses motivos, a equipe decidiu mudar a proposta do projeto para ter como alvo um dataset mais fácil de manipular, mas ainda tendo o foco em segmentação semântica e detecção de tumores.

Assim, o novo dataset escolhido foi o ”Breast Ultrasound Images Dataset” [4] encontrado no Kaggle. Ele é um conjunto que possui imagens de ultrassom mamário de cerca de 600 mulheres, entre seus 25 e 75 anos. As ultrassonografias foram coletadas durante o ano de 2018, para o treinamento de modelos de *machine learning*, em tarefas de classificação

e segmentação de tumores mamários. O conjunto de dados consiste em 780 imagens, no formato PNG com um tamanho médio de imagem de 500 × 500 pixels. As imagens são categorizadas em três classes, que são: normal, *benign* (benigno) e *malignant* (maligno). As 3 categorias são:

- Normal: Células normais mamárias do corpo humano.
- Benign: Classe representativa de tumores cancerígenos, porém benignas ao corpo humano.
- Malignant: Câncer maléfico que se espalha para outros sistemas do corpo.

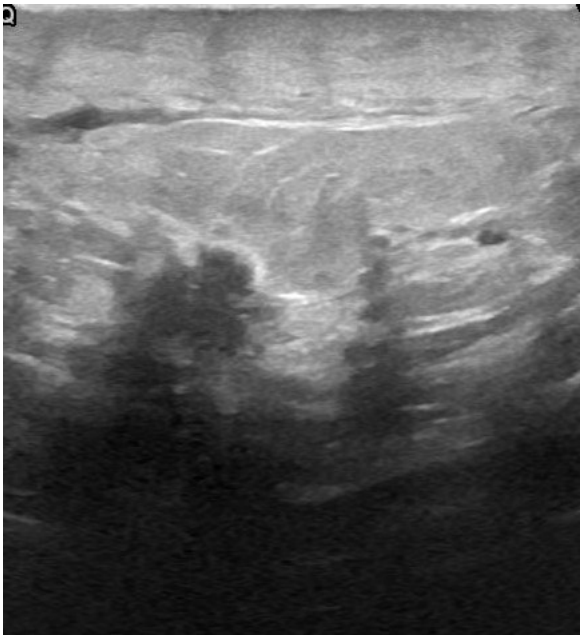

Figura 1. Exemplo de imagem com tumor malígno

No novo banco de dados, foram feitas 3 principais formas de manipulação de suas imagens. A primeira foi o tratamento de imagens que possuiam mais de uma máscara correspondente, significando a aparição de mais de um tumor. A junção dessas máscaras foram feitas de forma manual, pois houveram apenas 17 ocasiões desse tipo, pela plataforma online plataforma online Photopea [5], uma ferramenta de edição de imagens similar ao Adobe Photoshop, utilizando o modo de mistura "Tela" [6]. A segunda manipulação foi alterar a resolução das imagens do dataset para 256x256. A escolha original para a resolução das imagens era 500x500, por ser a resolução média das imagens do dataset [4], levando a uma deformação menor das imagens ao reescalá-las; porém as capacidades computacionais que a equipe conseguiu providenciar para o projeto não foram suficientes para suportar imagens com essas dimensões. Usar a transformação de corte de imagem também foi considerada, pois a reescala de imagens pode gerar inconsistências ou perda de detalhes da imagem original [7], mas a equipe notou que várias imagens possuiam máscaras com tamanhos muito variados, em proporção com as dimensões da imagem, e uma vez que os exemplos fossem cortados, as máscaras poderiam ocupar o corte final totalmente, assim como poderiam ser completamente ocultadas; em ambos os casos, a equipe julgou que prejudicaria o aprendizado dos modelos. Por último, exemplos de imagens da classe "Normal"foram excluídas do treinamento, pois os modelos tendiam a prever todas as imagens como normais, sem a segmentação de turmores; assim diminuindo para 647 a quantidade de exemplos do dataset, onde 80% desses formaram o dataset de treinamento, e os 20% restantes, o dataset de teste e validação.

B. Implementação das redes

Inicialmente, foram propostas 3 redes neurais convolucionais para serem comparadas: a U-Net, a R2U-Net e a FastFCN. Porém, a tentativa de comparar as redes U-Net e FastFCN com a R2U-Net, se mostrou como outro problema durante o desenvolvimento dos modelos, pois esta é uma variação da rede neural U-Net que apresenta não só conexões residuais, mas também é uma rede neural recorrente. Durante a criação da proposta de projeto, o grupo não tinha consciência de que nesta situação, usar uma rede recorrente não tem sentido, considerando que não se trabalharia com sequências de imagens. Por esse motivo, após um melhor estudo do grupo sobre redes neurais recorrentes, também foi decidido que a rede neural R2U-Net seria substituída pela rede Residual U-Net (ResU-Net), que é uma variação da U-Net, assim como a anterior, porém sem recorrência, apresentado apenas as conexões residuais.

Dessa forma, foram implementadas as 3 seguintes redes: U-Net, ResU-Net e FastFCN. Todas as redes usaram os mesmos datasets de treinamento

A U-Net e a ResU-Net foram feitas a partir da mesma classe "Unet", onde foi adicionado um parâmetro booleano na criação da classe chamado "residual", com o valor padrão igual a "False". Quando "residual"for verdadeiro, a rede U-Net declarada terá suas conexões residuais ativadas nos blocos convolucionais (blocos compostos de duas camadas convolucionais com tamanho de *kernel* 3x3 e *padding* igual a 1) presentes no encoder e decoder da rede U-Net, assim tornando-a em uma ResU-Net. Apesar do posicionamento exato das conexões residuais variar entre implementações do modelo, elas sempre conectam o início, ou o meio, do bloco convolucional-residual ao seu fim e igualmente, no nosso modelo, as conexões realizam uma convolução 1x1 com o tensor de entrada do bloco e adicionam o resultado ao tensor que passou pelas camadas convolucionais, antes da 2ª ReLU, a função de ativação usada dentro dos blocos convolucionais.

Já a FastFCN é constituída por duas partes principais, a *Joint Pyramid Upsampling* (JPU) e a cabeça da FastFCN. Para nossa implementação, a FastFCN original foi adaptada para se adequar às dimensões escolhidas, assim como a rede neural pré-treinada usada na JPU. No *paper* original da FastFCN, a rede pré-treinada utilizada foi a Resnet50, porém o espaço consumido pela importação da rede ocupava grande parte do espaço oferecido pelo Google Colab, levando o grupo a optar pelo uso de uma rede VGG16 pré-treinada, ao invés da Resnet50. Além disso, também foi necessário modificar a VGG16 pré-treinada para suportar imagens com apenas

um canal (L) que o banco de dados oferece, ao invés dos tradicionais três canais (RGB).

C. Treinamento

Inicialmente, foram testadas várias *loss functions* para alcançar os melhores resultados possíveis e melhor comparar as diferentes CNNs implementadas. As funções de perda testadas foram:

- *Binary-Cross-Entropy Loss* (*BCE Loss*): A *Cross-entropy*, ou Entropia-cruzada, pode ser definida como a medida da diferença entre duas distribuições de probabilidade para uma variável aleatória ou conjunto de eventos. É muito usada como função de perda em tarefas de segmentação. A *BCE Loss* é a função adaptada para problemas de segmentação semântica binária.
- *BCE with Dice coefficient Loss* (*DiceBCE Loss*): Uma variação da *BCE Loss* que soma seu resultado com o resultado de outra função, a *Dice Loss*, tentando aproveitar o melhor de ambas perdas. A *Dice Loss* é uma função de perda que foi adaptada para o uso do coeficiente *Dice*, que também é usado para avaliação de modelos de segmentação.
- *Focal Tversky Loss*: A *Tversky Loss* é uma função de perda que usa uma variação do coeficiente *Dice* em seu cálculo; nela as classificações *pixelwise* falso-positivas e falso-negativas recebem pesos maiores. A versão "*Focal*"dessa função é uma variante da anterior que dá um foco maior para exemplos com uma região de interesse menor.

Após os testes, a pior *loss function* foi a *BCE Loss*; já com a *Focal Tversky Loss* e a *DiceBCE Loss*, apresentaram resultados satisfatórios e diferenças entre predições de arquiteturas iguais mas com *loss functions* diferentes foram pequenas. Portanto o grupo decidiu usar as duas últimas funções para analisar e comparar as redes neurais. Da comparação do resultado a partir do uso das duas funções de perda, perecebemos uma tendência nas imagens de tomada de maior risco nas previsões, e avaliação de falsos positivos, da *Focal Tversky Loss*. Este resultado poderia ser utilizado num problema real, em situações que se é preciso que o modelo identifique mais que o necessário para uma posterior análise visual de um especialista, no caso do nosso problema um médico.

Além disso, todas as redes foram treinadas com o tamanho de *batch* igual a 8, e com o número de épocas igual a 40, apesar de testes iniciais terem sido feitos com apenas 5 épocas, devido às limitações da plataforma utilizada, Google Colab.

V. Resultados

Como mencionado, o objetivo do projeto é analisar a eficiência dos modelos escolhidos comparativamente a fim de identificar pontos positivos e negativos na utilização de cada um deles. Dessa forma, definimos métricas de avaliação para que haja uma forma objetiva de realizar análises correlacionais alheias a subjetividade.

1) Pixel Accuracy: Calcula a proporção dos pixels corretamente classificados em relação ao total de pixels na tela. Métrica para avaliação geral do desempenho. Dada a natureza das imagens a segmentação binária permitiu uma taxa alta de acurácia de pixels classificados, dessa forma, a taxa calculada do desempenho dos modelos aprensentou valores parecidos tanto entre modelos, quanto entre funções de perda. O valor encontrado variou entre 0.94 e 0.96 para os modelos. Sendo:

Modelos CNN	Funções de perda	
	Focal Tversky	**DiceBCE**
U-Net	0.9513	0.9679
ResU-Net	0.9404	0.9503
FastFCN	0.9595	0.9608

2) Intersection over Union: O Intersection over Union (IoU) mede a diferença entre a máscara de segmentação predita e a máscara esperada.

Modelos CNN	Funções de perda	
	Focal Tversky	**DiceBCE**
U-Net	0.6242	0.7231
ResU-Net	0.5429	0.6465
FastFCN	0.7033	0.7041

3) Avaliação de observação: Avaliação da qualidade da segmentação por meio da comparação com a imagem original. Dessa forma, será possível enxergar as falhas na segmentação do modelo avaliado, e assim entender as capacidades e limitações do reconhecimento de formas e objetos na cena. Da análise das imagens com suas máscaras preditas pudemos observar uma adequação satisfatória nos 3 modelos avaliados. Observando composições que montamos do IoU evidenciamos uma boa assertividade, apesar de existirem também recorrentemente a segmentação de falsos positivos, o que de certa forma era esperado dada a natureza complexa das imagens médicas.

4) Outros: Também foi observado que, usando uma GPU V100 providenciada pelo Google Colab, e com os parâmetros já mencionados de tamanho dos *batches* e número de épocas, ambas redes U-Net e ResU-Net eram treinadas em cerca de 13 minutos, enquanto as redes FastFCN eram treinadas em 8 minutos, com ambas as *loss functions*.

VI. Conclusão

A ResU-Net foi a que obteve o pior resultado entre as 3 redes avaliadas. O motivo disso, acreditamos, foi a natureza das imagens que eram preto e branco e extremamente abstratas, tornando a parte residual da rede menos efetiva. Além disso, a baixa profundidade de nossa rede também pode ter tido influência, dado que uma rede pouco profunda não precisaria de guardar informações das imagens antes da decodificação; porém não temos como comprovar esta suposição devido às limitações do Google Colab e do *hardware* dos membros da equipe, que não conseguem lidar com redes computacionalmente mais exigentes. A U-net treinada com DiceBCE como função de perda foi a que obteve o melhor resultado na métrica Intersecttion over Union, e avaliamos ela como sendo a melhor para a tarefa de segmentação de imagens médicas como as do

problema que nos propusemos a trabalhar. Sua assertividade alinhada com a facilidade de acesso e aplicação dela foram os pontos que destacamos como principais para a escolha dela como vencedora. Já a JPU da FastFCN adiciona uma complexidade maior à sua arquitetura, que não se converteu, em nossas avaliações, em uma melhora significativa da assertividade na resolução do problema, apenas em tempo de treinamento. A dificuldade de entender e aplicar esta rede é um empecílio que enfrentamos e que podem atrapalhar outros que se proponham a utilizar o modelo para solução de alguma tarefa. Na questão do desempenho, entretanto, ficamos satisfeitos sendo ela a rede que apresentou o desempenho mais equilibrado quando testada de diferentes funções de perda. Acreditamos que tanto U-Net quanto FastFCN são redes que resolvem de forma satisfatória o problema de segmentação de imagens de câncer de mama, com um favorecimento à U-Net por sua acessibilidade.

REFERÊNCIAS

[1] "Breast Cancer - Metastatic: Statistics" - https://www.cancer.net/cancer-types/breast-cancer-metastatic/statistics.

[2] "BACH: Breast Cancer Histology Images" - https://www.kaggle.com/datasets/truthisneverlinear/bach-breast-cancer-histology-images

[3] "Adobe Photoshop and SVS files - Pathology" - https://pathology.med.umich.edu/digital-pathology/adobe-photoshop-and-svs-files

[4] AL-DHABYANI, Walid; GOMAA, Mohammed; KHALED, Hussien; FAHMY, Aly. Dataset of breast ultrasound images. Cairo, 21 nov. 2019. DOI: https://doi.org/10.1016/j.dib.2019.104863. Disponível em: https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset/

[5] "Photopea" - https://www.photopea.com/

[6] "Modos de mesclagem no Adobe Photoshop" - https://helpx.adobe.com/pt/photoshop/using/blending-modes.html

[7] "Chapter 14.9. Semantic Segmentation and the Dataset — Dive into Deep Learning" - https://d2l.ai/chapter_computer-vision/semantic-segmentation-and-dataset.html

[8] R. Olaf et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation"2015, doi: https://doi.org/10.48550/arXiv.1505.04597.

[9] Wu, Huikai, et al. "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation."arXiv preprint arXiv:1903.11816 (2019)..