Semantic Segmentation for People Detection on Beach Images

Leonardo de A. Monte[1], **Emília G. Oliveira**[1], **Filipe R. Cordeiro**[1] **e Valmir Macario**[1]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
Caixa Postal 15.064 – 91.501-970 – Recife – PE – Brazil

{leonardo.amonte, emilia.galdino, filipe.rolim, valmir.macario}@ufrpe.br

Abstract. *Our work analyses a set of semantic segmentation methods applied to detect people on beach images, as part of an automatic track system to monitor bathers to prevent trespassing the boundaries of the safe region for swimming. In our analysis, we compared the semantic segmentation networks U-net, X-net, LinkNet and Unet++ with the pretrained backbones VGG-16 and VGG-19. We built our own dataset, composed of 300 images. The models were evaluated bound box-wise using F-score metrics. Our findings show that X-Net achieved the best value for F-score, with 90.89%, while Linknet was faster than the other networks, with no significant statistical difference in the F-score values.*

Resumo. *Nosso trabalho compara um conjunto de redes de segmentação semântica aplicados na detecção de pessoas em imagens de praia, como parte de um sistema de rastreamento automático para evitar que banhistas ultrapassem a região segura do mar. Em nossa análise, comparamos as redes de segmentação U-net, X-net, Linknet, e Unet++ usando os backbones pré-treinados VGG-16 e VGG-19. Nós propomos nossa própria base de imagens, composta de 300 imagens. Os modelos foram avaliados utilizando a métrica F-score. Nossos resultados mostraram que a Linknet obteve o melhor valor de F-score, com 90.89%, enquanto a Linknet foi mais rápida que as outras redes, sem diferença estatística significativa.*

1. Introdução

Ambientes de praia podem apresentar diversos riscos para os banhistas, como afogamentos e ataques de tubarão, o que torna o monitoramento destes locais uma tarefa importante para a prevenção de acidentes. De acordo com as estatísticas divulgadas pelo relatório do Comitê Estadual de Monitoramento de Incidentes com Tubarões (CEMIT) [CEMIT 2021] do estado de Pernambuco, Brasil, ocorreram 62 incidentes com tubarão no estado, dos quais 24 ocorreram na praia de Boa Viagem, no Recife. Devido ao histórico de acidentes com tubarão, por toda a costa da praia de Boa Viagem existem placas sinalizando os banhistas para que eles permaneçam dentro dos limites das barreiras de coral, ou em locais com água até a altura da cintura, devido ao risco de ataque de tubarão. Entretanto, mesmo com a sinalização, alguns banhistas ultrapassam as barreiras seguras da praia, tornando-se assim mais vulneráveis a acidentes.

O monitoramento utilizando câmeras de segurança vem sendo auxiliado por sistemas de visão computacional para identificar situações de risco [Chen et al. 2020]. Por conta da necessidade de monitoramento constante de algumas dessas áreas para prevenir

situações de perigo, um sistema automatizado seria uma medida efetiva de controle de risco. Dentro desse contexto, os maiores desafios envolvendo o problema detecção de pessoas em ambiente de praia são a variação de iluminação nas imagens, oclusão parcial e o posicionamento distante das câmeras [Chevtchenko et al. 2018]. Outra limitação se encontra na aquisição de exemplos positivos, no nosso caso, imagens de pessoas dentro da água: por conta da sinalização de aviso, a maior parte dos banhistas evitam o mar nestas áreas. A segmentação semântica tem ganhado notoriedade em diversas tarefas de visão computacional, como carros autônomos [Miclea and Nedevschi 2019], interação homem-máquina [Wong et al. 2017], segmentação de texto escrito à mão [Jo et al. 2020], e outras aplicações. A arquitetura da segmentação semântica consiste em dois módulos: *encoder* e *decoder*. O módulo *encoder* extrai as características, também conhecidas como *embeddings*, a partir da imagem, e é criado a partir de modelos clássicos de *Deep Learning*, como VGG [Simonyan and Zisserman 2015], ResNets [He et al. 2015], também chamadas de modelo de *backbone*. O módulo *decoder* então é aplicado ao modelo de *backbone* de forma a restaurar a resolução das características (ex: U-net, LinkNet).

Este trabalho compara as seguintes redes de segmentação semântica no problema de detecção de pessoas em imagens de praia: U-net [Ronneberger et al. 2015], Xnet [Bullock et al. 2018], LinkNet [Chaurasia and Culurciello 2017] e Unet++ [Zhou et al. 2018]. Para os *backbones*, foram utilizadas a VGG-16 e VGG-19. Cada uma das redes de segmentação semântica foi selecionada de acordo com os seguintes critérios: ter sido criada para obter boa performance em bases de dados reduzidas, ou ter um tempo de treinamento e inferência rápidos em comparação com outras redes de segmentação semântica. Estes critérios são relevantes para o problema pois a detecção precisa ser realizada em tempo real e devido ao fato da base de dados criada ser composta de 300 imagens, que é pequena em comparação com outras bases de dados de segmentação. As performances das redes neurais foram avaliadas utilizando as métricas *F-score*, *precision*, *recall* apenas nas áreas de mar da imagem, uma vez que o principal objetivo é evitar acidentes na faixa de mar, e portanto as pessoas que estão na área de areia da imagem não são considerados. Para avaliar os melhores resultados, o teste de Wilcoxon [Stapor 2017] foi utilizado na métrica *F-score*. De forma a selecionar a meta-arquitetura mais adequada para o problema, também foi calculada a média de *frames* por segundo para cada rede de segmentação semântica. Nosso trabalho é motivado pelo recente sucesso dos algoritmos de segmentação semântica para detecção de pessoas, incluindo abordagens que detectam objetos pequenos em tempo real. É apresentado um estudo de redes de segmentação semântica que partem da premissa de obter bons resultados em bases de dados pequenas, ou de obter tempos de treinamento e de inferência rápidos, aplicadas na tarefa de detecção de banhistas em imagens de praia.

2. Trabalhos Relacionados

2.1. Detecção de pessoas em imagens de praia

A tarefa de detecção de pessoas em cenários de praia foi abordada na literatura como pode ser visto no trabalho de [Green et al. 2005], no qual os autores desenvolveram um sistema de detecção de pessoas utilizando uma base de dados formada por imagens de pessoas e não-pessoas (objetos que poderiam ser classificados erroneamente como pessoas). Os objetos foram segmentados utilizando a busca em largura e o detector de bordas de *Canny*. Os resultados utilizando um *Multi-layer Perceptron* foram de 91%

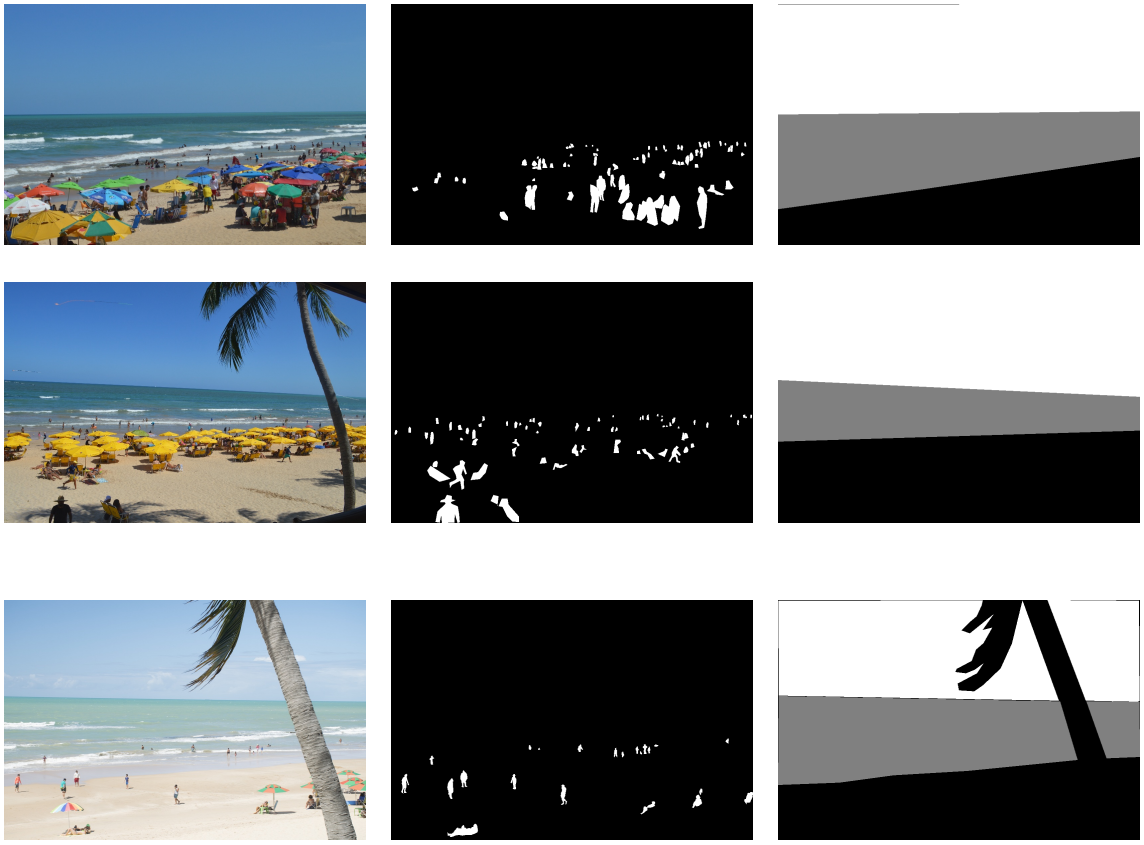

Figura 1. Exemplos de imagens de base de dados. Na primeira coluna estão as imagens originais. Na segunda coluna estão as imagens dos *labels* de pessoas. Na terceira coluna estão as imagens separadas pelas áreas de mar, céu e areia.

dos verdadeiros positivos e de 13% dos falso positivos. Seguindo a mesma ideia, em [Luna da Silva et al. 2017] os autores propuseram um sistema de detecção de pessoas em imagens de praia utilizando um conjunto diferente de descritores de características e de classificadores. Os autores utilizaram uma base de dados formada por imagens de pessoas e não-pessoas tiradas em ambientes de praia. A melhor taxa de reconhecimento foi de 90.31% obtida usando a técnica de PCA, uma combinação dos descritores HOG e LBP e Máquinas de Vetores de Suporte utilizando a função radial. Ambos trabalhos apresentam o problema de não conseguir detectar mais de uma pessoa por vez, assim como a sua localização exata na imagem de praia, devido ao fato de ambos utilizarem imagens de pessoas e objetos previamente cortadas, um problema que poderia ter sido resolvido utilizando a abordagem de segmentação semântica. O trabalho de [Chevtchenko et al. 2018] utiliza meta-arquiteturas de *Deep Learning Faster* R-CNN, R-FCN e SSD combinadas com algoritmos de classificação pré-treinados na base de dados COCO. Os autores descobriram que os detectores SSD foram uma ordem de magnitude mais rápidos do que as meta-arquiteturas R-FCN, e *Faster* R-CNN, porém tiveram dificuldades para detectar objetos distantes. Por outro lado, os autores verificaram que *Faster* R-CNN com a Resnet 101 obteve resultados de detecção significantemente melhores, mas com uma taxa de 5.6 *frames* por segundo usando uma placa de vídeo GTX 1080 Ti. Apesar da acurácia das soluções propostas para detecção de pessoas em imagens de praia, os resultados destes

algoritmos são resultados de *bounding-boxes*, perdendo a localização precisa das pessoas e necessitando um maior tempo de execução e de detecção.

2.2. Segmentação Semântica

Algoritmos de segmentação semântica têm sido utilizados em trabalhos recentes da literatura com o objetivo de obter resultados mais promissores em imagens classificadas pixel a pixel de forma mais natural, reduzindo significantemente a complexidade e o custo computacional do treinamento, além de serem fiéis às localizações das pessoas em tempo real [Siam et al. 2018, Liu et al. 2019b, Yang et al. 2020]. O trabalho de [Siam et al. 2018] fez um estudo do *trade-off* entre acurácia e custo computacional, utilizando uma combinação de meta-arquiteturas com diferentes *backbones* e *encoders* para segmentar objetos da base de dados *Citiscapes*. No trabalho de [Liu et al. 2019b] são buscados pontos centrais nos quais existem pedestres, e é proposto um novo método chamado de *Center and Scale Prediction* (CSP). O método proposto possui dois componentes, o módulo de extração de características e o módulo de detecção de cabeças. O algoritmo CSP se tornou o novo estado da arte em performance em dois *benchmarks* de detecção de pedestres, *Cityscapes* e *Caltech*. A *Narrow Deep Network* (NDNet) foi proposta em [Yang et al. 2020], que usou uma estrutura de convolução separável do tipo *bottleneck*, modificando a estrutura da rede neural totalmente convolucional8 (FCN8) [Long et al. 2015] com uma fusão do *score* aprendido, e aumento de objetos pequenos para identificar objetos menores na base de dados *Cityscapes*. Diferente dos outros trabalhos de segmentação semântica, o trabalho a ser proposto neste artigo é um estudo de redes de segmentação com *backbones* pré-treinados e não pré-treinados, testados no problema de detecção de pessoas em imagens de praia, que ainda não foi avaliado com esta abordagem.

3. *Background*

3.1. Segmentação Semântica

A segmentação semântica pode ser vista como um processo de classificação a nível de pixel, no qual cada imagem de entrada é classificada pixel a pixel de acordo com a classe correspondente. O problema de classificação a nível de pixel pode ser entendido da seguinte forma: Encontre uma forma de atribuir a uma classe do conjunto de classes $L = \{l_1, l_2, \ldots, l_k\}$ para cada um dos pixels de uma imagem 2D, com dimensões de W x $H = N$ pixels. Cada classe l representa uma classe diferente ou objeto, que pode ser por exemplo avião, carro, semáforo, ou no caso deste trabalho, banhista. O espaço do conjunto de classes possui k possíveis classes, que são comumente estendidas para $k + 1$ sendo l_0 a classe de *background* ou a classe vazia [Garcia-Garcia et al. 2017].

Muitas das técnicas de segmentação semântica que formam o estado da arte seguem a estrutura básica da Rede neural totalmente conectada proposta em [Long et al. 2015]. A ideia principal é aproveitar redes de *Deep Learning* de classificação a partir da troca das camadas totalmente conectadas por um mapa de probabilidades das mesmas dimensões da imagem de entrada, indicando a probabilidade de cada pixel pertencer a cada uma das classes pré-definidas. Estes mapas passam pelo processo de *upsampling* usando operações de deconvolução, como por exemplo a interpolação bilinear, para produzir saídas densas em classes por pixel [Garcia-Garcia et al. 2017, Noh et al. 2015].

3.2. Abordagens relacionadas

Nesta seção serão descritas algumas das técnicas do estado da arte de segmentação semântica que foram utilizadas neste trabalho.

3.2.1. U-net

A U-net [Ronneberger et al. 2015] é uma Rede Neural Convolucional para segmentação semântica, que foi criada para segmentação de imagens médicas. Devido à pouca disponibilidade de imagens médicas, esta rede neural foi criada com o propósito de obter resultados eficientes utilizando uma pequena quantidade de dados. A rede neural recebeu este nome de acordo com o formato da sua arquitetura em U, que ocorre devido à sua configuração de camadas, na qual a primeira metade é chamada de *contracting path* e a segunda metade é chamada de *expansive path*. O *contracting path* é a primeira etapa da rede neural que realiza uma série de convoluções seguidas por etapas de *maxpooling* nos dados, de forma a extrair informação em diferentes níveis da imagem. O *expansive path* é a segunda parte da rede, na qual são realizadas convoluções seguidas de *upsamplings*, para recuperar o tamanho original da imagem e extrair informações úteis durante o processo.

3.2.2. Xnet

A X-net [Bullock et al. 2018] é uma Rede Neural Convolucional que, tal como a U-net, também foi criada para segmentação de imagens médicas, mais especificamente para imagens de raio-x, e que é aplicável ao caso de bases de dados pequenas. No primeiro quarto das transformações, a rede neural realiza convoluções seguidas de *maxpoolings*. Depois desse processo, é realizado o processo de *upsampling* e é extraída a localização precisa da informação na imagem. Este processo é repetido mais uma vez, e ao final é aplicada uma camada de ativação para a classificação pixel a pixel da imagem segmentada.

3.2.3. LinkNet

A LinkNet [Chaurasia and Culurciello 2017] é uma Rede Neural Convolucional para segmentação semântica criada para obter resultados similares aos do estado da arte, porém com um custo computacional baixo e sem necessariamente aumentar a quantidade de parâmetros da rede. O módulo *encoder* consiste em convoluções seguidas de reduções de dimensionalidade da imagem por um fator dois, enquanto que o *decoder* consiste em convoluções seguidas de aumento na dimensionalidade de imagem, também por um fator dois, de forma que seja possível extrair informações em diferentes escalas e recuperar as dimensões originais para a classificação pixel a pixel.

3.2.4. Unet++

A Unet++ [Zhou et al. 2018] é uma Rede Neural Convolucional para segmentação semântica criada a partir da arquitetura da U-net, consistindo na aplicação de um *con-*

tracting path e de um *expansive path* assim como a U-net. No *contracting path* são realizadas convoluções seguidas de *maxpoolings*, de forma que seja possível extrair diferentes tipos de informação da imagem, ao mesmo tempo reduzindo o seu tamanho original. No *expansive path* são realizadas convoluções seguidas de *upsamplings* de forma que seja possível extrair informações da mesma forma que no *contracting path*, enquanto restaura a imagem para o seu tamanho original. Entre cada um dos estágios do *contracting path* e do *expansive path* é realizada uma série de convoluções densas, de forma que a rede possa propagar a ativação de uma parte da rede para outra com uma menor diferença semântica. Cada camada possui uma função de *skip connection* que consiste em propagar as camadas de ativação do *contracting path* para o *expansive path*.

4. Avaliação Experimental

4.1. *Encoder* e *Decoder*

Os algoritmos de segmentação semântica: U-net, Unet++, X-net, e LinkNet, foram avaliados com diferentes *backbones*: VGG-16 e VGG-19 [Simonyan and Zisserman 2015]. Foram avaliados os *backbones* treinados utilizando a base de dados proposta e também um modelo pré-treinado utilizando *transfer learning* para a base de dados proposta. Para os *backbones* pré-treinados, foram utilizados os pesos dos modelos treinados na base de dados *ImageNet* [Deng et al. 2009], que contém 1000 classes de objetos, incluindo pessoas em diferentes níveis de escala e oclusão.

4.2. Base de dados

A base de dados de imagens consiste em fotos que foram tiradas a partir de postos de salva-vidas na praia de Boa Viagem, Recife-Brasil. A base proposta consiste em 300 imagens, pertencentes a uma das duas possíveis classes, pessoa e *background*, na qual tudo o que não é uma pessoa na imagem é classificado como *background*. A base de dados contém um total de 14.023 pessoas, cuja segmentação foi realizada manualmente, utilizando a plataforma *LabelMe* [Wada 2016]. As pessoas são apresentadas nas imagens em diferentes níveis de oclusão, nos quais objetos como guarda-sóis ocultam partes do corpo das pessoas. Além da oclusão por objetos, algumas pessoas na imagem estão parcialmente imersas na água, algumas vezes sendo visível apenas uma parte do corpo, como a cabeça por exemplo, fazendo da detecção de pessoas nos cenários de praia um problema ainda mais desafiador. A figura 1 mostra três exemplos de imagens da base de dados proposta. As imagens da primeira coluna são as imagens originais, na segunda coluna estão as imagens segmentadas das pessoas, e na terceira coluna estão imagens classificadas de acordo com áreas de mar, céu e areia.

4.3. Método de avaliação

Como a preocupação principal neste trabalho é detectar pessoas na imagem, foi criada uma métrica para avaliar o quão bem a rede de segmentação semântica realiza a detecção baseada na métrica do desafio PASCAL *Visual Object Classes* (VOC) [Everingham et al. 2010], descrita a seguir. Para cada pessoa existe um *ground-truth bounding box*, que é usado para calcular os verdadeiros positivos, falsos negativos e falsos positivos. A partir da lista das coordenadas dos *ground-truth bounding box*, é extraída a região delimitada pelas coordenadas tanto da imagem segmentada manualmente, quanto do resultado da rede de segmentação semântica. A próxima etapa é o cálculo da interseção entre as

duas imagens binárias resultantes da etapa anterior. Os verdadeiros positivos são computados caso a área de interseção seja ao menos do mesmo tamanho que uma fração da área da segmentação manual da região, de acordo com um limite pré-estabelecido. Caso contrário, é considerado como um falso negativo. Para o cálculo de falsos positivos é necessário primeiramente calcular a área de segmentação média das pessoas na parte de mar nas imagens da base de dados. Esta etapa ocorre devido ao fato de que os falsos positivos são estimados a partir da área residual após a remoção de todos os *pixels* pertencentes aos verdadeiros positivos e falsos negativos. A área residual é dividida pela área de segmentação média anterior para estimar o número de falsos positivos. Por fim, é usado o número de verdadeiros positivos, falsos positivos e falsos negativos para calcular as métricas de *precision*, *recall*, *F-score*.

4.4. Seleção de Parâmetros

Para cada rede neural um conjunto de hiper-parâmetros foi selecionado baseado inicialmente na implementação original do artigo, e então os hiper-parâmetros foram refinados de acordo com os melhores resultados nos testes experimentais. Os itens a seguir descrevem cada hiper-parâmetro utilizado nas redes de segmentação semântica.

- **Taxa de aprendizado:** Para a *U-net* a taxa de aprendizado utilizada foi de 0.001, para a *X-net* foi usada a de 0.0001, para a *Linknet* foi de 0.001, e para a *Unet++* foi de 0.0003. Todas as redes de segmentação semântica utilizaram *learning rate decay* com o valor de *decay* como sendo 0.0001.
- ***Backbone*:** Cada uma das redes de segmentação semântica foi avaliada utilizando uma arquitetura CNN como *backbone*. Os *backbones* utilizados nos experimentos foram a *VGG-16* e a *VGG-19* [Simonyan and Zisserman 2015].
- **Função de perda:** A função de perda usada nas redes de segmentação semântica foi o coeficiente de *Jaccard*, também conhecido como interseção sobre união [Berman et al. 2018].
- **Otimizador:** Para a *U-net*, *Xnet*, e *Unet++* o otimizador usado foi o Adam, e para a *LinkNet* foi usado o *RMSprop*, de acordo com os resultados experimentais.
- **Número de épocas:** Para os experimentos foi selecionado o número de 200 para o número de épocas, também foi utilizada a técnica de *early stopping* de forma que o número de épocas fosse suspenso em caso de não haver melhoras na performance por 25 épocas, para evitar *overfitting*.
- **Inicialização de pesos:** Cada rede de segmentação semântica foi avaliada utilizando duas formas de inicialização de pesos. A primeira forma foi utilizando os pesos da rede neural pré-treinada na base de dados *imagenet* [Krizhevsky et al. 2012]. A segunda forma foi inicializar os pesos utilizando a inicialização Uniforme *Glorot* [Hanin and Rolnick 2018], que é a inicialização padrão da biblioteca *Keras* [Chollet et al. 2015], e portanto não utiliza pesos pré-treinados.

5. Resultados

Para a avaliação experimental foi usada a técnica de validação cruzada *5-fold* [Raschka 2020], executada 6 vezes com diferentes conjuntos de imagens da base de dados, e todos os experimentos foram conduzidos no mesmo computador com um processador i7. Foram avaliadas duas configurações para as redes neurais: com e sem *encoders* pré-treinados. Para os modelos pré-treinados, foram utilizados os pesos dos modelos pré-treinados na

Figura 2. Exemplo de imagem de entrada.

Figura 3. Exemplo de segmentação da *Linknet*.

Figura 4. Exemplo de segmentação da *U-net*.

Figura 5. Exemplo de imagem ouro.

Figura 6. Exemplo de segmentação da *X-net*.

Figura 7. Exemplo de segmentação da *Unet++*.

Figura 8. Comparação qualitativa dos resultados produzidos pelas redes de segmentação semântica.

base de dados *ImageNet*, enquanto que para os modelos não pré-treinados, foi usada a inicialização uniforme *Glorot*. Para cada resultado são apresentados os valores de média e desvio padrão, respectivamente. As tabelas 1 e 2 mostram os resultados da avaliação utilizando a métrica desenvolvida no projeto.

Sobre a avaliação das redes não pré-treinadas, é possível observar na tabela 1 que a *U-net* com o *backbone VGG-19* atingiu os melhores resultados e a *Linknet* obteve os piores resultados, com 18.08% para o *F-score*. As redes pré-treinadas alcançaram melhores resultados em comparação com as redes não pré-treinadas, das quais a *X-net* usando o *backbone VGG-16* obteve os melhores resultados e a *Linknet* diferentemente do experimento com a rede não pré-treinada, obteve resultados acima de 90%, como pode ser visto na tabela 2. Os resultados demonstram que a *Xnet* com o *backbone VGG-16* alcançou o melhor resultado nas métricas de *recall* e *F-score*. Na métrica *precision* os melhores resultados foram obtidos pela *Unet* com o *backbone VGG-19*. Dado o problema de detecção de banhistas, a métrica de *recall* se mostra importante deviso às situações que podem colocar em risco à vida dos banhistas, caso algum banhista não seja detectado. Além da métrica de *recall*, outra métrica importante é a métrica *precision*, pois enquanto a detecção de todos os banhistas é crítica, a rede neural também deve não gerar um grande número de alarmes falsos. Os valores de *recall* das redes e a análise qualitativa da segmentação representada na figura, mostraram que as redes foram capazes de segmentar a maioria das pessoas na faixa de mar, mas alguns pixels foram perdidos durante o processo, o que fez com que o resultado da métrica *precision* diminuísse. No geral, as redes pré-treinadas

Tabela 1. Média e desvio padrão dos resultados das redes **não** pré-treinadas nas métricas de avaliação baseadas em *bouding box*.

Modelo	***Precision***	***Recall***	***F-score***
Unet-VGG16	87.38(**0.06**)	93.84(**0.08**)	90.10(**0.05**)
Xnet-VGG16	86.89(**0.06**)	92.48(0.09)	89.13(**0.05**)
Linknet-VGG16	17.43(0.34)	18.95(0.38)	18.08(0.36)
Unet++-VGG16	70.31(0.35)	71.12(0.37)	70.07(0.35)
Unet-VGG19	87.48(**0.06**)	**94.23**(**0.08**)	**90.37**(**0.05**)
Xnet-VGG19	86.73(**0.06**)	91.04(0.11)	88.20(0.06)
Linknet-VGG19	26.05(0.39)	28.69(0.44)	27.21(0.41)
Unet++-VGG19	**87.49**(**0.06**)	91.63(0.10)	88.95(0.06)

Tabela 2. Média e desvio padrão para os resultados dos experimentos com redes pré-treinadas usando a métrica de avaliação baseada em *bouding box*.

Modelo	***Precision***	***Recall***	***F-score***
Unet-VGG16	87.60(**0.05**)	94.50(0.08)	90.56(0.05)
Xnet-VGG16	86.94(**0.05**)	**95.80**(**0.06**)	**90.89**(**0.03**)
Linknet-VGG16	87.66(**0.05**)	94.81(0.08)	90.76(0.05)
Unet++-VGG16	86.71(0.06)	94.62(0.07)	90.16(0.04)
Unet-VGG19	**87.92**(**0.05**)	94.24(0.08)	90.63(0.05)
Xnet-VGG19	86.90(0.06)	95.62(**0.06**)	90.77(0.04)
Linknet-VGG19	87.52(**0.05**)	94.37(0.08)	90.45(0.05)
Unet++-VGG19	87.01(**0.05**)	94.89(0.07)	90.47(0.04)

obtiveram melhores resultados em comparação com as redes não pré-treinadas, com a maior diferença sendo entre a *Linknet*, na qual a rede não pré-treinada não atingiu mais do que 20% na métrica de *F-score*, enquanto que a rede pré-treinada obteve mais de 90% na mesma métrica. O teste de *Wilcoxon* com confiança de 5% foi aplicado de forma a avaliar os melhores resultados para a métrica de *F-score*. As melhores configurações para a métrica de *F-score*, em cada rede (XnetVGG16, LinknetVGG16, UnetVGG19 and Unet++VGG19) foram comparadas entre si. Todas as comparações entre a rede *Unet++* e as outras redes obtiveram *p-value* menor do que o valor de confiança, enquanto as outras redes obtiveram *p-value* maiores do que o valor de confiança, rejeitando portanto a hipótese de que os valores da métrica de *F-score* das redes comparadas fossem provenientes de distribuições diferentes. Outro aspecto interessante deste projeto é a expectativa de que as redes performem a detecção de pessoas em tempo real. Um sumário comparativo entre os tempos de inferência das redes de segmentação semântica é essencial para escolher a melhor rede para o problema em questão. A figura 9 mostra a comparação entre as médias de *frames* por segundo para o tempo de inferência das redes de segmentação semântica no conjunto de testes.

A figura 9 mostra que a *Linknet* obteve a melhor média de *frames* por segundo, enquanto a *Xnet* obteve os piores resultados em ambos *backbones* utilizados. Como os melhores resultados para a métrica de *F-score* pela *Xnet*, *Unet* e *Linknet* não mostraram diferenças estatísticas pelo teste de *Wilcoxon*, e a velocidade de detecção da *Linknet* é 2 vezes mais rápida do que a *X-net* e mais rápida que a *U-net*, a configuração de rede e

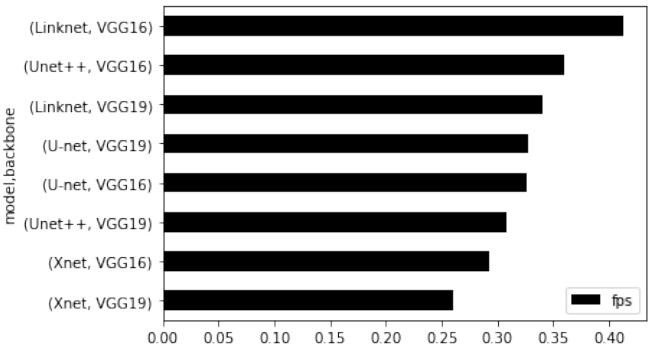

Figura 9. Comparação entre as velocidades de detecção usando CPU.

backbone mais viável de ser utilizada para o problema de detecção de pessoas em imagens de praia foi a *Linknet* com o *backbone* da *VGG-16*.

6. Conclusões e trabalhos futuros

Este trabalho analisou quatro redes de segmentação semântica aplicadas ao problema de detecção de pessoas em ambiente de praia. As detecções são desafiadoras devido às variações nas condições de tempo, posicionamento distante de câmera e oclusão parcial. Para cada rede neural, dois *backbones* diferentes foram utilizados, *VGG-16* e *VGG-19*, e todas as redes com pesos pré-treinados na base de dados *Imagenet* obtiveram resultados melhores em comparação com as redes com os pesos inicializados com a inicialização uniforme *Glorot*. As redes *Linknet*, *Xnet* e *Unet* mostraram boas performances em áreas de praia, de forma que elas poderiam ser futuramente aplicadas para ambientes similares, com oclusão parcial e grandes variações de iluminação nas imagens. Além disso, os modelos foram também comparados em termos de *frames* por segundo, com a *Linknet* usando o *backbone VGG-16* obtendo tempo de inferência 2 vezes mais rápido em comparação com a *Xnet* e também mais rápido do que a *Unet*. Como trabalhos futuros, novas técnicas como *Conditional Random Fields* CRF) [Zhou et al. 2016] podem ser utilizadas para refinar as saídas das redes de segmentação, além de *Neural Architecture Search* (NAS) [Liu et al. 2019a] para buscar por outras variações de arquiteturas que otimizem a performance dos algoritmos de segmentação.

Referências

[Berman et al. 2018] Berman, M., Triki, A. R., and Blaschko, M. B. (2018). The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks.

[Bullock et al. 2018] Bullock, J., Cuesta-Lázaro, C., and Quera-Bofarull, A. (2018). Xnet: A convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets. *CoRR*, abs/1812.00548.

[CEMIT 2021] CEMIT (2021). Statistics of shark incidents in the state of pernambuco-brazil.

[Chaurasia and Culurciello 2017] Chaurasia, A. and Culurciello, E. (2017). Linknet: Exploiting encoder representations for efficient semantic segmentation. *CoRR*, abs/1707.03718.

[Chen et al. 2020] Chen, C., Surette, R., and Shah, M. (2020). Automated monitoring for security camera networks: promise from computer vision labs. *Security Journal*.

[Chevtchenko et al. 2018] Chevtchenko, S., Vale, R., Cordeiro, F., and Macario, V. (2018). Deep learning for people detection on beach images. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 218–223.

[Chollet et al. 2015] Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

[Deng et al. 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[Everingham et al. 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

[Garcia-Garcia et al. 2017] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*.

[Green et al. 2005] Green, S., Blumenstein, M., Browne, M., and Tomlinson, R. (2005). The detection and quantification of persons in cluttered beach scenes using neural network-based classification. In *Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05)*, pages 303–308.

[Hanin and Rolnick 2018] Hanin, B. and Rolnick, D. (2018). How to start training: The effect of initialization and architecture.

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

[Jo et al. 2020] Jo, J., Koo, H. I., Soh, J. W., and Cho, N. I. (2020). Handwritten text segmentation via end-to-end learning of convolutional neural networks. *Multimedia Tools and Applications*, 79(43):32137–32150.

[Krizhevsky et al. 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

[Liu et al. 2019a] Liu, C., Chen, L.-C., Schroff, F., Adam, H., Hua, W., Yuille, A. L., and Fei-Fei, L. (2019a). Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Liu et al. 2019b] Liu, W., Liao, S., Ren, W., Hu, W., and Yu, Y. (2019b). High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5187–5196.

[Long et al. 2015] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

[Luna da Silva et al. 2017] Luna da Silva, R., Chevtchenko, S., Alves de Moura, A., Rolim Cordeiro, F., and Macario, V. (2017). Detecting people from beach images. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 636–643.

[Miclea and Nedevschi 2019] Miclea, V.-C. and Nedevschi, S. (2019). Real-time semantic segmentation-based stereo reconstruction. *IEEE Transactions on Intelligent Transportation Systems*, 21(4):1514–1524.

[Noh et al. 2015] Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528.

[Raschka 2020] Raschka, S. (2020). Model evaluation, model selection, and algorithm selection in machine learning.

[Ronneberger et al. 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

[Siam et al. 2018] Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., and Jagersand, M. (2018). Rtseg: Real-time semantic segmentation comparative study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1603–1607. IEEE.

[Simonyan and Zisserman 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Stapor 2017] Stapor, K. (2017). Evaluating and comparing classifiers: Review, some recommendations and limitations. pages 12–21.

[Wada 2016] Wada, K. (2016). labelme: Image Polygonal Annotation with Python. `https://github.com/wkentaro/labelme`.

[Wong et al. 2017] Wong, J. M., Kee, V., Le, T., Wagner, S., Mariottini, G.-L., Schneider, A., Hamilton, L., Chipalkatty, R., Hebert, M., Johnson, D. M., et al. (2017). Segicp: Integrated deep semantic segmentation and pose estimation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5784–5789. IEEE.

[Yang et al. 2020] Yang, Z., Yu, H., Feng, M., Sun, W., Lin, X., Sun, M., Mao, Z.-H., and Mian, A. (2020). Small object augmentation of urban scenes for real-time semantic segmentation. *IEEE Transactions on Image Processing*, 29:5175–5190.

[Zhou et al. 2016] Zhou, H., Jun Zhang, Jun Lei, Shuohao Li, and Dan Tu (2016). Image semantic segmentation based on fcn-crf model. In *2016 International Conference on Image, Vision and Computing (ICIVC)*, pages 9–14.

[Zhou et al. 2018] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer.