Segmentação Semântica de Cenas Urbanas com Cityscapes: Uma Análise Comparativa de Modelos de Aprendizagem Profunda

1st Camila Vieira
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
cbv2@cin.ufpe.br

***Abstract*—A segmentação semântica é crucial para uma interpretação mais precisa e detalhada do conteúdo visual, contribuindo significativamente para o entendimento visual. Nesse contexto, o conjunto de dados Cityscapes, composto por mais de 20.000 imagens de cenas urbanas de diversas cidades, com algumas apresentando anotações de alta qualidade e a maioria com anotações mais simplificadas, serve como base para uma análise comparativa entre modelos de aprendizado profundo, incluindo U-Net, SegNet, FCN, DeepLab e PSPNet. Esse campo continua a presenciar avanços constantes, devido às inúmeras aplicações práticas, como o desenvolvimento de veículos autônomos, planejamento urbano, monitoramento ambiental e realidade aumentada e virtual.**

***Index Terms*—Segmentação Semântica, Compreensão de Cena, Cityscapes Dataset, Cenas Urbanas, Modelos de Aprendizagem Profunda, Análise Comparativa**

I. Introdução

A compreensão visual é um dos principais desafios enfrentados na área de visão computacional, envolvendo a identificação e classificação de objetos, o entendimento das relações espaciais, a inferência de contexto e a interpretação completa do cenário como um todo. Uma máquina com essa capacidade avançada seria capaz de passar no Teste de Turing, respondendo corretamente a qualquer questionamento relacionado a uma determinada imagem e imitando o comportamento humano.

Para alcançar esse nível de inteligência artificial, a estrutura de dados requer uma complexidade suficiente para abranger todo o conhecimento relevante sobre a cena observada. É essencial conter informações detalhadas sobre os elementos presentes e permitir inferências sobre qualquer aspecto necessário, mesmo diante da ampla diversidade visual que pode ser encontrada. Sendo preciso determinar tanto a essência da imagem, quanto a quantidade de rachaduras que pode haver em uma parede no segundo plano.

Dentre os elementos fundamentais para tornar essa compreensão visual possível, destaca-se a segmentação semântica, que atribui rótulos semânticos a cada pixel da imagem, classificando-os em categorias específicas. A detecção e localização de objetos identificam itens específicos presentes na cena, enquanto o reconhecimento de atributos descreve características adicionais, como direção do movimento, cor e aparência. O conhecimento de contexto assemelha-se à capacidade humana de compreender um cenário com base no que não está diretamente presente. Além disso, o raciocínio espacial estabelece relações entre objetos e suas posições relativas, enquanto o rastreamento e detecção de movimento acompanham a posição e o comportamento dos objetos ao longo do tempo.

A segmentação semântica desempenha um papel crucial ao fornecer uma interpretação mais precisa e detalhada do conteúdo visual. Sendo assim, é de grande relevância investigar técnicas e modelos de aprendizagem profunda que possam aprimorar o desempenho dessa tarefa. Nesse contexto, o uso do dataset Cityscapes, um conjunto de dados abrangente contendo cenas urbanas de mais de 50 cidades, oferece a oportunidade de realizar uma análise comparativa das arquiteturas de redes neurais convolucionais mais promissoras da área, como U-Net, SegNet, FCN, DeepLab e PSPNet.

O aperfeiçoamento da segmentação semântica, e, por conseguinte, da compreensão de cena, proporciona avanços significativos em diversas aplicações práticas. Isso inclui melhorias nos sistemas de percepção e tomada de decisão de veículos autônomos, possibilitando uma identificação mais precisa de obstáculos, semáforos e faixas de pedestres. Além disso, possibilita a identificação de áreas com maior tráfego e a análise do fluxo de pedestres para contribuir com o planejamento urbano inteligente. Acompanhamento da qualidade do ar e das áreas verdes permite o monitoramento ambiental, enquanto a detecção de atividades suspeitas e comportamentos anormais é fundamental para análises de segurança. Por fim, a integração mais realista de elementos virtuais ao ambiente urbano impulsiona aplicações em realidade aumentada e virtual, facilitando a navegação e manipulação de objetos. Em geral, tais avanços contribuem para o desenvolvimento de tecnologias mais inteligentes e seguras para as cidades do futuro.

II. Objetivo

Propõe-se a realização uma análise comparativa entre diferentes modelos de aprendizagem profunda aplicados à tarefa de

segmentação semântica de cenas urbanas usando o conjunto de dados Cityscapes. Busca-se avaliar o desempenho e a capacidade de generalização de modelos amplamente utilizados, incluindo U-Net, SegNet, FCN, DeepLab e PSPNet, a fim de compreender suas limitações e identificar os pontos fortes de cada arquitetura.

Além disso, pretende-se fornecer uma análise completa e detalhada dos resultados obtidos, utilizando métricas quantitativas, como Intersection over Union (IoU) e Mean IoU (mIoU), bem como análise qualitativa para visualizar exemplos de segmentações geradas por cada modelo. A partir dessas avaliações, evidencia-se as capacidades e desafios enfrentados pelas arquiteturas em cenários urbanos complexos. Assim, orientando futuros aprimoramentos nos sistemas de segmentação semântica e contribuindo para o desenvolvimento de soluções mais eficientes e precisas em aplicações práticas, como veículos autônomos, mapeamento urbano e análise de tráfego.

III. Justificativa

A segmentação semântica é uma tarefa de grande relevância na área de visão computacional devido às suas inúmeras aplicações práticas em diversos campos. A compreensão detalhada do conteúdo visual em cenários urbanos é fundamental para o desenvolvimento de tecnologias avançadas, como veículos autônomos, planejamento urbano inteligente, monitoramento ambiental, segurança pública e realidade aumentada. Nesse contexto, a análise comparativa de modelos de aprendizagem profunda utilizando o conjunto de dados Cityscapes é essencial para identificar as arquiteturas mais eficientes e precisas para essa tarefa. Busca-se aprender sobre o desempenho e a capacidade de generalização de cada modelo, contribuindo para o aprimoramento de sistemas de segmentação semântica e avançando o estado-da-arte em compreensão visual. Além disso, a utilização do Cityscapes, um conjunto de dados abrangente, realista e desafiador, permite avaliar o desempenho dos modelos em cenários variados, impulsionando o desenvolvimento e avaliação de algoritmos e modelos para soluções tecnológicas mais inteligentes.

IV. Metodologia

A. Dataset

O conjunto de dados selecionado para este estudo é o Cityscapes, criado em 2016 pelo Grupo de Pesquisa de Autonomia Visual do Instituto Max Planck de Informática e do Instituto de Ciência da Computação da Universidade de Tübingen, ambos localizados na Alemanha. O Cityscapes é um conjunto abrangente de imagens de alta resolução de cidades em ambientes urbanos, capturadas a partir de gravações de cenas de tráfego de transporte público em mais de 50 cidades grandes da Alemanha, França e Suíça, ao longo de vários meses, variando condições, como iluminação, variação climática e densidade populacional.

O principal objetivo do dataset é fornecer dados realistas e desafiadores de cenários urbanos do mundo real que possam ser utilizados em projetos de pesquisa e competições, incentivando o desenvolvimento e a avaliação de algoritmos e modelos avançados para a tarefa de segmentação semântica. O Cityscapes é amplamente reconhecido e utilizado na comunidade de pesquisa em visão computacional devido à sua riqueza de classes, detalhes e realismo.

O conjunto de dados Cityscapes é composto por um total de 5.000 imagens de alta resolução, com uma resolução típica de 2048x1024 pixels. As imagens estão divididas em três principais conjuntos: treinamento, validação e teste, garantindo uma divisão representativa e aleatória dos dados. Cada imagem é acompanhada de uma máscara de segmentação no mesmo formato, onde cada pixel é rotulado com um valor numérico indicativo da classe semântica à qual pertence, totalizando mais de trinta classes possíveis.

As anotações fornecidas no Cityscapes são detalhadas e de alta qualidade. Cada classe é rotulada com precisão, resultando em uma segmentação de alta resolução e precisa. Além disso, o conjunto de dados disponibiliza informações adicionais, como anotações de instâncias individuais para veículos e pessoas, bem como metadados, como tempo, dados de profundidade ou vários quadros de vídeos. Outros pesquisadores contribuíram disponibilizando anotações de caixas delimitadoras de pessoas e imagens aumentadas com efeitos de neblina e chuva.

A política de rotulamento seguida no Cityscapes dita que objetos em primeiro plano nunca devem ter buracos, mesmo que o plano de fundo seja visível através do objeto. O mesmo deve ser seguido para regiões com muita sobreposição de duas ou mais classes. Quanto às anotações de instâncias individuais, quando a fronteira entre duas instâncias não for clara, toda a região será demarcada como ”grupo”.

O conjunto de dados Cityscapes também inclui várias categorias de competição, como rotulagem semântica em nível de pixel, rotulagem semântica em nível de instância, rotulagem semântica panóptica e detecção de veículo 3D.

Além das imagens com anotações finas, o download do Cityscapes também disponibiliza imagens com anotações grosseiras e arquivos adicionais que complementam as imagens e anotações de segmentação. Esses arquivos adicionais incluem informações de disparidade para as imagens, parâmetros da câmera para cada imagem e máscaras de segmentação específicas para veículos.

O uso dos dados do Cityscapes exige a concordância com os termos de uso, que proíbem o uso comercial e são destinados apenas para fins de pesquisa. Os pesquisadores devem citar adequadamente o projeto ao utilizar o conjunto de dados. É importante destacar que não há garantia de que o conjunto de dados esteja tecnicamente ou matematicamente correto, podendo conter erros.

Ao selecionar o conjunto de dados Cityscapes para este estudo, visa-se utilizar um conjunto realista e diversificado de cenas urbanas, permitindo uma análise comparativa sólida dos modelos de aprendizagem profunda para a tarefa de segmentação semântica.

Fig. 1. Exemplo de máscara com anotação fina.

Fig. 2. Cidade inclusas no Cityscapes.

TABLE I
CLASSES INCLUSAS NO CITYSCAPES

Grupo	**Classes**
flat	road · sidewalk · parking · rail track
human	person · rider
vehicle	car · truck · bus · on rails · motorcycle · bicycle · caravan · trailer
construction	building · wall · fence · guard rail · bridge · tunnel
object	pole · pole group · traffic sign · traffic light
nature	vegetation · terrain
sky	sky
void	ground · dynamic · static

B. Modelos e Arquiteturas

Os modelos de aprendizagem profunda selecionados para a análise comparativa de segmentação semântica de cenas urbanas usando o conjunto de dados Cityscapes foram U-Net, SegNet, FCN, DeepLab e PSPNet. As arquiteturas escolhidas foram selecionados com base em sua popularidade e eficácia comprovada em tarefas de segmentação semântica. O objetivo é avaliar o desempenho e a capacidade de generalização de cada modelo em cenários urbanos complexos e variados.

1) U-Net: A U-Net é um dos modelos mais populares para segmentação semântica devido à sua arquitetura encoder-decoder com conexões residuais. A estrutura em formato de ”U” permite que informações de níveis de resolução mais altos sejam transmitidas para a fase de decodificação, o que facilita a recuperação de detalhes importantes na etapa de segmentação.

2) SegNet: O modelo SegNet também utiliza uma arquitetura encoder-decoder, mas com uma abordagem mais leve em relação à U-Net. Essa arquitetura foi projetada especificamente para a tarefa de segmentação semântica e busca um equilíbrio entre desempenho e eficiência computacional.

3) FCN: O modelo FCN (Fully Convolutional Network) é uma arquitetura que transforma redes neurais convolucionais profundas em modelos para segmentação semântica. Ele permite a preservação da resolução espacial durante o processo de convolução, capturando informações contextuais em várias escalas.

4) DeepLab: O modelo DeepLab utiliza dilatações nas camadas convolucionais para aumentar o campo receptivo, melhorando a precisão da segmentação. Essa abordagem é particularmente eficaz para a captura de informações contextuais em uma área maior da imagem.

5) PSPNet: A arquitetura PSPNet ((Pyramid Scene Parsing Network) utiliza pirâmides de pooling para capturar informações contextuais em diferentes escalas, possibilitando uma compreensão mais abrangente da cena.

Cada modelo será implementado utilizando a biblioteca PyTorch, e seu treinamento e avaliação serão conduzidos no ambiente do Google Colab, aproveitando o acesso gratuito a GPUs para acelerar o processo de aprendizagem. A divisão dos dados em conjuntos de treinamento, validação e teste seguirá as diretrizes padrão do conjunto de dados Cityscapes, garantindo uma avaliação justa e representativa do desempenho de cada modelo.

C. Métricas de Avaliação

As métricas de avaliação são fundamentais para analisar e comparar o desempenho dos modelos de segmentação semântica de cenas urbanas. Pretende-se avaliar os resultados obtidos pelos modelos por métricas quantitativas, como Precision, Pixel Accuracy, IoU, mIoU e IoU por Classe, e por análise qualitativa, então realizar análise de erros e comparação dos tempos de execução. Com isso, objetiva-se identificar os pontos fortes e fracos de cada arquitetura e aprender como aprimorar o sistema

1) Precision: A precisão avalia a proporção de verdadeiros positivos em relação a todos os exemplos classificados como positivos. Em outras palavras, mede a capacidade do modelo de classificar corretamente os exemplos positivos.

$$Precision = \frac{TP}{(TP + FP)} \tag{1}$$

2) Pixel Accuracy: A acurácia de pixels mede a proporção de pixels corretamente classificados (verdadeiros positivos e verdadeiros negativos) em relação ao total de pixels da imagem. É útil para avaliar o desempenho geral da segmentação.

$$PixelAccuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{2}$$

3) Intersection over Union (IoU): O IoU, também conhecido como Jaccard Index, é uma métrica comumente utilizada para medir a sobreposição entre a máscara de segmentação predita e a máscara de verdade absoluta em cada classe. O IoU é calculado como a razão entre a interseção e a união das duas máscaras, conforme a fórmula:

$$IoU = \frac{TP}{(TP + FP + FN)} \tag{3}$$

onde TP representa o número de pixels verdadeiro positivos, FP é o número de pixels falso positivos e FN é o número de pixels falso negativos, todos referentes à classe em questão. O IoU varia de 0 a 1, onde 1 indica uma segmentação perfeita e 0 indica nenhuma sobreposição entre as máscaras.

4) Mean IoU (mIoU): Para avaliar o desempenho geral dos modelos, calcula-se a média do IoU para todas as classes presentes no conjunto de dados Cityscapes. O mIoU é uma métrica importante para comparar a capacidade de segmentação dos modelos em diferentes classes. Um valor mais próximo de 1 indica uma segmentação precisa e abrangente em todas as classes, enquanto valores mais baixos indicam falhas em segmentar certas classes.

5) IoU por Classe: Além do mIoU global, também apresenta-se os valores de IoU individuais para cada classe presente no conjunto de dados Cityscapes. Isso permite uma análise detalhada do desempenho dos modelos em cada categoria de interesse. Identificar quais classes são segmentadas com alta precisão e quais classes apresentam desafios para os modelos é essencial para compreender as capacidades e limitações dos modelos em cenários urbanos.

6) Análise Qualitativa: Além das métricas quantitativas, realiza-se uma análise qualitativa dos resultados, visualizando exemplos de segmentações geradas pelos modelos. Assim, apresentando imagens do conjunto de teste juntamente com as respectivas máscaras de segmentação para ilustrar como os modelos lidam com diferentes classes e cenários urbanos. Essa análise visual permite identificar falhas específicas dos modelos, bem como pontos fortes em suas segmentações.

A combinação de métricas quantitativas e qualitativas proporciona uma avaliação completa e detalhada do desempenho dos modelos de aprendizagem profunda para a segmentação semântica de cenas urbanas. Essas métricas são essenciais para compreender a capacidade de generalização dos modelos, identificar suas limitações e orientar futuros aprimoramentos nos sistemas de segmentação semântica para aplicações práticas em veículos autônomos, mapeamento urbano e análise de tráfego.

D. Ambiente de Experimentação

O ambiente de experimentação selecionado para este estudo é o "Google Colab" ou "Colaboratory", uma poderosa ferramenta em nuvem que permite a execução de código Python diretamente pelo navegador, sem a necessidade de configurações complexas. O Google Colab oferece uma série de benefícios, incluindo a facilidade de compartilhamento, a capacidade de combinar código, texto, imagens e outros recursos em um único ambiente colaborativo.

Uma das principais vantagens do Google Colab é o acesso gratuito a Unidades de Processamento Gráfico (GPUs) e Unidades de Processamento Tensorial (TPUs) fornecidos pelo Google. Esses recursos computacionais acelerados são especialmente valiosos para tarefas intensivas em processamento, como segmentação semântica de cenas urbanas usando modelos de aprendizagem profunda. Ao utilizar GPUs e TPUs na nuvem, os pesquisadores podem se beneficiar do alto poder computacional sem restrições do hardware da sua máquina, impulsionando o desempenho e a velocidade dos experimentos.

No entanto, é importante observar que a versão gratuita do Google Colab apresenta limitações que precisam ser consideradas ao realizar experimentos mais complexos e extensos. O tempo máximo de execução contínua é limitado a doze horas, após isso o ambiente será reinicializado automaticamente. Para contornar essa limitação, é aconselhável salvar os resultados intermediários.

Outra consideração importante é a disponibilidade de recursos de hardware. Embora o Google Colab ofereça acesso a GPUs e TPUs gratuitamente, a disponibilidade desses recursos pode variar, e não há garantia de que uma GPU ou TPU estará sempre disponível quando necessário. Para contornar a situação, os pesquisadores podem tentar executar experimentos em horários de menor demanda, quando a probabilidade de obter acesso a recursos de hardware é maior.

No que diz respeito ao armazenamento e acesso aos dados, é necessário fazer o upload dos conjuntos de dados para o Google Drive, o que pode ser um processo demorado, especialmente para grandes volumes de dados. Recomenda-se a organização cuidadosa dos dados e o uso eficiente de técnicas de carregamento e pré-processamento para minimizar o tempo de espera durante o experimento. Além disso, é importante salvar os dados relevantes no Drive para evitar perda de informações após reinicialização do ambiente.

Para escapar dessas limitações e maximizar a eficiência dos experimentos, uma abordagem estratégica pode ser adotada. Isso inclui a otimização do código para aproveitar ao máximo os recursos disponíveis, o uso de técnicas de pré-processamento para reduzir o tamanho dos dados durante a fase de carregamento e a implementação de estratégias de salvamento e backup para preservar resultados intermediários importantes.

Em casos de estouro da memória RAM no ambiente do Google Colab, é possível adotar estratégias para lidar com esse problema, como o uso de amostragem dos dados ou a redução do tamanho dos lotes (batch size) durante o treinamento dos

modelos de segmentação. Dessa forma, mesmo com recursos limitados de memória, é viável continuar a experimentação usando apenas parte dos dados ou processando-os em lotes menores.

Adicionalmente, o PyTorch, uma das principais bibliotecas de aprendizagem profunda utilizadas no Google Colab, oferece uma função que permite o salvamento de etapas intermediárias durante o treinamento dos modelos. Essa funcionalidade, conhecida como "checkpointing", possibilita salvar os pesos do modelo, otimizador e outros parâmetros em pontos específicos do treinamento. Com isso, em caso de reinicialização da sessão, é possível retomar o treinamento a partir do ponto salvo, evitando a perda de progresso e acelerando o processo de treinamento.

Apesar dessas limitações, as medidas adequadas para contorná-la tornam o Google Colab gratuito um ambiente poderoso e acessível para realizar análises comparativas de modelos de aprendizagem profunda para segmentação semântica de cenas urbanas usando o conjunto de dados Cityscapes.

V. Conclusões e Discussões

A segmentação semântica é uma tarefa fundamental na área de visão computacional, com aplicações que vão desde a condução autônoma até o reconhecimento de objetos em imagens médicas. Neste projeto, tinha-se como objetivo realizar uma análise comparativa de cinco modelos de segmentação semântica (U-Net, SegNet, FCN, DeepLab e PSPNet) utilizando o conjunto de dados Cityscape. No entanto, houveram limitações significativas de recursos computacionais que impossibilitaram a execução dos experimentos planejados.

A principal limitação prevista foi a falta de capacidade de processamento na GPU e, por isso, diversas formas de contornar possíveis problemas foram estudas. Embora já esperava-se que o treinamento de modelos de segmentação semântica fosse intensivo em termos computacionais, a situação foi agravada quando a RAM do sistema não suportou as demandas do treinamento. Mesmo após a implementação de várias estratégias de otimização, não conseguimos contornar esse problema, o que prejudicou a execução dos experimentos.

Essa situação destaca a importância de considerar cuidadosamente os recursos computacionais disponíveis ao planejar projetos de visão computacional. Além disso, ressalta a necessidade de otimização de algoritmos e estratégias para treinamento de modelos em ambientes com recursos limitados, uma área de pesquisa cada vez mais relevante à medida que a demanda por modelos de alto desempenho continua a crescer.

Para versões futuras deste projeto, é essencial que sejam realizadas otimizações de parâmetros e estratégias para permitir a execução bem-sucedida dos experimentos. Isso pode incluir o uso de técnicas de treinamento mais eficientes, como a transferência de aprendizado, o ajuste fino de hiperparâmetros e o uso de arquiteturas de rede mais leves. A busca pela melhor versão de cada modelo é crucial para garantir uma análise comparativa precisa e útil.

Apesar das limitações encontradas, é fundamental ressaltar a importância deste trabalho. A segmentação semântica desempenha um papel crucial em diversas aplicações do mundo real, como a melhoria da segurança no trânsito, a automação de tarefas industriais e a assistência médica avançada. A pesquisa e a avaliação de diferentes modelos de segmentação semântica são passos essenciais para avançar nessa área e permitir o desenvolvimento de sistemas mais robustos e precisos. Embora este projeto tenha enfrentado desafios significativos, ele representa um primeiro passo valioso em direção a futuros avanços na segmentação semântica e em direção a soluções mais eficazes e acessíveis para problemas do mundo real.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth e B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding", arXiv:1604.01685v2 [cs.CV], 7 de abril de 2016.

[2] J. Xiao, B. C. Russell, J. Hays, K. A. Ehinger, A. Oliva e A. Torralba, "Basic Level Scene Understanding: From Labels to Structure and Beyond", em Massachusetts Institute of Technology, University of Washington e Brown University.

[3] New Jersey Institute of Technology. "Scene Understanding Introduction." Disponível em: https://pantelis.github.io/cs677/docs/common/lectures/scene-understanding/scene-understanding-intro/. Acesso em: 2023-07-21.

[4] Cityscapes. "Cityscapes Dataset". Disponível em: https://www.cityscapes-dataset.com/.

[5] Colab. "Olá, este é o Colaboratory". Disponível em: https://colab.research.google.com/notebooks/welcome.ipynb.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol. 9351, 2015, pp. 234-241.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," arXiv:1511.00561 [cs.CV], Nov. 2, 2015. [Online]. Disponível em: https://arxiv.org/abs/1511.00561

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in CVPR 2017, Computer Vision and Pattern Recognition (cs.CV), arXiv:1612.01105 [cs.CV], 2016.

[9] L. P. Vieira, M. Simões, R. P. D. Ferraz, e J. A. Ribeiro, "Deep learning e segmentação semântica de imagens para diagnóstico de níveis de degradação de pastagem," Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1153427/1/Deep-learning-e-segmentacao-semantica-de-imagens-2023.pdf.