Métodos de *Deep Learning* aplicados à Segmentação Semântica de Imagens para Percepção de Veículos Autônomos

Gabriel Toffanetto França da Rocha
Laboratório de Mobilidade Autônoma – LMA
Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas

Campinas, Brasil
g289320@dac.unicamp.br

***Abstract*—**
- Veículos autonomos
- Visão computacional
- Percepção do ambiente
- Segmentação Semântica de Imagem
- Métodos Vanilla
- Métodos Deep Learning
- Necessidades da aplicação
- Resultados
- Proximos passos (teste para obtenção do Perception grid)

***Index Terms*—Deep learning, Visão computacional, Segmentação Semântica de Imagem, Robótica móvel, Veículos autônomos**

I. Introdução

Veículos com capacidade de se guiarem de forma autônoma estão cada vez mais presentes no dia a dia da sociedade contemporânea, possibilitando que o motorista possa realizar outras atividades durante a navegação, ou que o mesmo seja assistido em caso de alguma falha humana do condutor. Para que o automóvel seja capaz de se mover por conta própria, o mesmo deve ser capaz de perceber o ambiente, e sensores como sonares, radares, LiDARs e câmeras podem ser utilizados para tal. Porém, a câmera se faz como uma solução mais viável economicamente, e como visto na literatura, apresenta soluções que contemplam os desafios da navegação autônoma de veículos em ambientes urbanos, como visto nos trabalhos de [7] e [15].

Para que um veículo autônomo possa entender o ambiente à sua volta, é necessário que ele saiba reconhecer as entidades que o compõem, como por exemplo: estrada, veículos, calçadas, pedestres e vegetação, para que assim, o mesmo saiba diferenciar área navegável de obstáculos [11]. Para isso, o emprego da técnica de segmentação semântica de imagens, onde cada pixel da imagem é classificado de acordo com a entidade do ambiente da qual ele faz parte [9]. A Fig. 1 mostra a aplicação da técnica de segmentação semântica fundida à informação de profundidade dada por uma câmera *stereo*, permitindo a obtenção de um *grid* de percepção dinâmica local (DLP), que projeta no plano 2D o ambiente contendo a detecção de múltiplos objetos para que o veículo consiga planejar seu caminho [17].

Existem métodos de processamento de imagens que realizam o mascaramento de cada entidade da imagem, porém a definição de qual é a classe de cada segmento se faz desafiadora, sendo anteriormente empregada a utilização de redes neurais artificiais (ANNs) para tal, como feito por [16]. Porém, devido à utilização de ANNs somente para a classificação final, era necessário muito pré-processamento para realização da segmentação semântica. Com o desenvolvimento das redes neurais profundas (DNNs), obteve-se métodos com poder suficiente para que, dada uma imagem bruta de entrada e uma imagem de referência (*ground truth*) segmentada para comparação, a rede profunda consegue aprender como realizar a segmentação da imagem do ambiente urbano, como nas várias arquiteturas mostradas por [13]. Com a popularização desses métodos, já existem diversos conjuntos de dados para treinamento das DNNs, como os [5] e [2], [3]. Existem também *datasets* que trazem cenas ainda mais desafiadoras, como o [12] que apresenta imagens urbanas durante noites chuvosas.

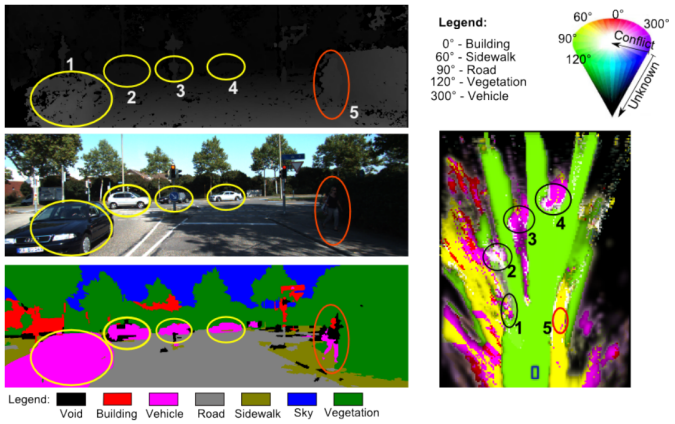

Fig. 1. DLP com ênfase na detecção múltipla de objetos móveis obtido com a fusão da imagem semanticamente segmentada e as informações de profundidade [17].

Dessa forma, a percepção do ambiente por meio de visão computacional se faz indispensável para o desenvolvimento dos veículos autônomos, e com isso, os métodos de *deep learning* se fazem uma grande ferramenta para conseguir-se reconhecer as entidades de uma cena urbana com robustez às variações de luz e reflexos, sendo assim uma solução a ser

explorada. Além do desempenho da segmentação semântica, o tempo demandado para tal operação também é vital, uma vez que durante a navegação, todos os módulos operam em tempo real, e a quantidade de *frames* segmentados por segundo é uma informação importante. Dessa forma, esse trabalho propõem a utilização de redes neurais profundas com diferentes arquiteturas para a segmentação semântica de imagens de cenas urbanas, utilizados *datasets* da literatura e para testes finais, utilizando imagens reais adquiridas pelo autor.

Este trabalho é dividido em seis partes, onde na Seção I é realizada a motivação e contextualização da pesquisa e na Seção II é apresentado o estado da arte, discorrendo sobre as soluções utilizadas atualmente. Com isso, a Seção III apresenta a metodologia a ser utilizada neste trabalho, seguida dos resultados obtidos na Seção IV e sua análise na Seção V. Por fim, são apresentadas as conclusões na Seção VI e as referências utilizadas.

II. Estado da arte

- Estratégias para operação em real time
- Artigos *Survey* [10]
- Tipos de redes utilizadas [1], [4], [6], [14], [18]–[20]
- Resultados estado-da-arte [13]
  - mIoU
  - FPS

III. Metodologia

- Arquiteturas escolhidas
- Redes escolhidas
- *Datasets* escolhidos
  - Proposta de utilizar dados coletados no *campus*
- Método de treinamento
- Métricas utilizadas
- *Frameworks* utilizados
- *Hardware* utilizado

Para a solução do problema foi escolhida a rede STDC1-50, que apresentou maior velocidade nos testes realizados pelos autores do *survey* [13], com um desempenho de segmentação considerável. [...]

A. Arquitetura

A rede STDC1 [6] é baseada na arquitetura de *two-branch network*, onde existe uma bifurcação da rede em dois ramos, um ligado à extração de informações de contexto e o outro aplicado para obtenção de informações espaciais da entrada. A informação de contexto demanda de uma maior profundidade, implementando mecanismos de atenção em dois níveis para extração dos atributos necessários para a identificação das regiões da imagem, como mostrado na Fig. 2(a).

Porém, o fato de se processar a informação de entrada em dois ramos, traz um inconveniente para o problema de segmentação em tempo real, que é justamente o tempo de inferência da rede. Desta forma, a rede aplica uma estratégia elegante ao utilizar no lugar de um ramo com novas camadas, uma *skip connection* do *Stage 3*, para o bloco de fusão [...]. Para forçar a captura das informações espaciais, a saída da *skip connection* é utilizada em uma tarefa auxiliar de detecção de bordas de cada região semântica da imagem, por meio de uma *detail head*. Tal procedimento é realizado por meio de treinamento supervisionado, onde com a aplicação de filtros laplacianos na imagem rótulo de segmentação semântica, se obtém a imagem rótulo com as bordas de cada segmento da imagem de entrada, que representam os detalhes à serem capturados pelo ramo espacial, e a perda é computada por meio da saída da *detail head* e o rótulo obtido. Reforça-se que esse procedimento é realizado apenas em treinamento, enquanto durante a inferência a rede neural irá utilizar as habilidades aprendidas por meio do gradiente sobre a perda atrelada à segmentação e à detecção de bordas, sem realizar explicitamente a extração das bordas de cada região da imagem.

Cada estágio é composto da forma ... E realiza convolução da forma...

O mecanismo de atenção é ...

Por fim, a fusão é feita ... e o upsampling...

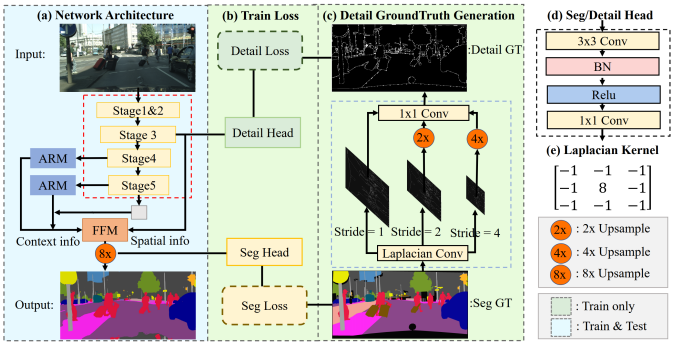

Fig. 2. Arquitetura completa da rede neural utilizada [6].

B. Dataset de treinamento e dados de teste

Dataset... tamanho... particionamento... resolução... classes...

C. Método de treinamento

Batch... epocas... iterações...

D. Métricas de avaliação

mIOU50... mIOU75...

E. Teste

IV. Resultados

- Para cada rede:
  - Métricas
  - Segmentação
    - Entrada
    - *Ground truth*
    - Saída
  - Tempo de treinamento

A. Métricas

B. Dados de teste

V. Análise dos Resultados

- Comparar as três redes:
  - mIoU
  - FPS
  - Dados segmentados do *dataset*
  - Dados segmentados coletados no *campus*
- Apontar relação custo *vs* desempenho de cada rede
- Considerar custos de treinamento

VI. Conclusões

- Retomar o problema inicial
- Destacar metodologia e os resultados que foram obtidos
- Comentar a análise dos resultados, mostrando que seria melhor para implementação
- Propor melhorias
- Propor validação de aplicação
- Listar proposta de aplicação dessa técnica
  - Perception grid

Agradecimentos

- Levy e Romis
- Giovani?

Referências

[1] Badrinarayanan, V., Kendall, A., and Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, Oct. 2016.

[2] Brostow, G. J., Fauqueur, J., and Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters 30*, 2 (Jan. 2009), 88–97.

[3] Brostow, G. J., Shotton, J., Fauqueur, J., and Cipolla, R. Segmentation and Recognition Using Structure from Motion Point Clouds. In *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., vol. 5302. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 44–57.

[4] Chao, P., Kao, C.-Y., Ruan, Y., Huang, C.-H., and Lin, Y.-L. HarDNet: A Low Memory Traffic Network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul, Korea (South), Oct. 2019), IEEE, pp. 3551–3560.

[5] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding, Apr. 2016.

[6] Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., and Wei, X. Rethinking BiSeNet For Real-time Semantic Segmentation, Apr. 2021.

[7] Garcia, O., Vitor, G. B., Ferreira, J. V., Meirelles, P. S., and De Miranda Neto, A. The VILMA intelligent vehicle: An architectural design for cooperative control between driver and automated system. *Journal of Modern Transportation 26*, 3 (Sept. 2018), 220–229.

[8] Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, second ed. O'Reilly, 2020.

[9] He, H.-J., Zheng, C., and Sun, D.-W. Image Segmentation Techniques. In *Computer Vision Technology for Food Quality Evaluation*. Elsevier, 2016, pp. 45–63.

[10] Janai, J., Güney, F., Behl, A., and Geiger, A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends® in Computer Graphics and Vision 12*, 1–3 (2020), 1–308.

[11] Jebamikyous, H.-H., and Kashef, R. Autonomous vehicles perception (AVP) using deep learning: Modeling, assessment, and challenges. *IEEE Access 10* (2022), 10523–10535.

[12] Jin, J., Fatemi, A., Lira, W., Yu, F., Leng, B., Ma, R., Mahdavi-Amiri, A., and Zhang, H. RaidaR: A Rich Annotated Image Dataset of Rainy Street Scenes, Oct. 2021.

[13] Papadeas, I., Tsochatzidis, L., Amanatiadis, A., and Pratikakis, I. Real-Time Semantic Image Segmentation with Deep Learning for Autonomous Driving: A Survey. *Applied Sciences 11*, 19 (Sept. 2021), 8802.

[14] Poudel, R. P. K., Bonde, U., Liwicki, S., and Zach, C. ContextNet: Exploring Context and Detail for Semantic Segmentation in Real-time, Nov. 2018.

[15] Vitor, G. B. *Urban environment and navigation using robotic vision: conception and implementation applied to autonomous vehicle = Percepção do ambiente urbano e navegação usando visão robótica: concepção e implementação aplicado à veículo autônomo*. Doutor em Engenharia Mecânica, Universidade Estadual de Campinas, Campinas, SP, Sept. 2014.

[16] Vitor, G. B., Lima, D. A., Victorino, A. C., and Janito V. Ferreira. A 2D/3D Vision Based Approach Applied to Road Detection in Urban Environments. In *2013 IEEE Intelligent Vehicles Symposium (IV)* (Gold Coast City, Australia, June 2013), IEEE, pp. 952–957.

[17] Vitor, G. B., Victorino, A. C., and Ferreira, J. V. Modeling evidential grids using semantic context information for dynamic scene perception. *Knowledge-Based Systems 215* (Mar. 2021), 106777.

[18] Wang, Y., Zhou, Q., and Wu, X. ESNet: An Efficient Symmetric Network for Real-time Semantic Segmentation, June 2019.

[19] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., and Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation, Apr. 2020.

[20] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation, Aug. 2018.