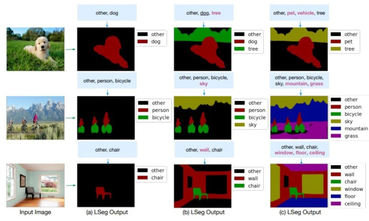



# 論文紹介：language-driven semantic segmentation

2022-11-14 16:00　AI技術開発部
AI, Tech

目次 [非表示]

1.はじめに
2.背景
3.モデル詳細
4.実験・評価
5.終わりに
6.引用

## はじめに

この記事では、language-driven semantic segmentationというタスクを行うモデルを提案した論文、Li et al. (2022)[1]を紹介したいと思います。language-driven semantic segmentationは、入力画像に対し、単語をいくつか与えてその単語に沿うようにsegmentationを実施する、というタスクになります。

## 背景

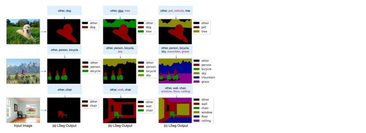
従来のsemantic segmentationは、画像の各ピクセルに対して分類するラベルが固定されていました。したがって、ラベルの数は訓練データセットによって決定され、数十から数百という単位になります。しかし、実際にはラベルとなる単語は膨大に存在し、ラベル単色をこのように限定することはsegmentationモデルの潜在的な認識能力を損ねている可能性があります。

ラベルの数が限定されている主な要因は、アノテーションのコストです。画像数、ラベル数が増えると、アノテーションに大変な労力が必要なのです。それを解決するために、zero-shot/few-shotのsegmentation手法が提案されています。

この論文では、semantic segmentationの柔軟性と汎用性を向上させるシンプルなzero-shotのアプローチ、「language-driven semantic segmentation」のLSegを提示しました。

LSegでは、ユーザはテスト時に任意の画像のラベルセットを自由に構成、並べ替えすることができ、そうして与えられたラベルセットに基づいて様々なsegmentation結果を柔軟に出力することができます。

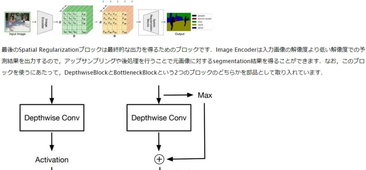
下図のstep1)のは、「対象物」と「その他」と指定されたラベルだけを識別しています。一方で、step2)のようにラベルが追加された場合、モデルは追加されたラベルに応じて対象物を柔軟にsegmentationしていることがわかります。




## モデル詳細

LSegは、テキストの各単語と画像の各ピクセルを埋め込み空間上にマッピングさせて学習します。

複数のラベルと画像が与えられた時、Text Encoderはラベルをベクトル空間に埋め込みます。Image Encoderは画像からピクセル単位の埋め込みを行い、各ピクセルの特徴を全ラベル埋め込みに対応させます。ここで、論文ではText Encoderとして学習済みCLIP[2]を使用し、Image EncoderとしてDPT（ViTをbackboneとしたモデル）[2]を使用しています。Image Encoderは、テキスト埋め込みと、そのピクセルの正解ラベルの最後にクロス埋め込みの相関を最大化するように学習します。Text Encoderは、概念的に近い単語は埋め込み空間上で近くなるように訓練されているので、与えられたデータセットのラベル単色で訓練するだけで、Text Encoderの柔軟性を画像認識部分でも享受できるようになります。




最後のSpatial Regularizationブロックは最終的な出力を得るためのブロックです。Image Encoderは入力画像の解像度よりも低い解像度での予測結果を出力するので、アップサンプリングや後処理を行うことで元画像に対するsegmentation結果を得ることができます。なお、このブロックを使うにあたって、DepthwiseBlockとBottleneckBlockというこの2つのブロックのどちらかを部品として取り入れています。

この2種類のSpatial Regularizationブロックについて、Ablation Studyを行っており、どちらも複数ブロックをスタックすることで改善することをわかっています。

| Block Type | Metric | # depth 0 | 1 | 2 | 4 |
|---|---|---|---|---|---|
| DepthwiseBlock | pixAcc [%] | 79.70 | 79.72 | **79.78** | 7.67 |
| | mIoU [%] | 37.83 | 39.19 | **39.45** | 0.18 |
| BottleneckBlock | pixAcc [%] | 79.70 | 79.64 | **79.70** | 79.68 |
| | mIoU [%] | 37.83 | 39.16 | **39.79** | 38.78 |

## 実験・評価

LSegはもともと推論に用いるラベルを学習中に一度も見ないzero-shotのモデルとして設計されていました。しかし、比較対象が十分でないため、zero-shotに加えfew-shotのモデル名いくつか用意し、LSegを比較しています。zero-shotのモデルとしては、ZS3Net[3]をベースラインとしています。

データセットはPASCAL-5iとCOCO-20iというものを使っています。これらはそれぞれPASCAL VOC 2012 (Everingham et al, 2015)[4] とCOCOデータセット (Lin et al, 2014)[5]から作成したfew-shot用のsegmentationデータセットです（赤字のはデータセットのフォールドを表します）。

LSegは従来のZS3Netを土台る構築であり、few-shotのモデルとも同等構築することがわかります。

LSegのバックボーンをResNet (He et al.,2016)[6]からViT (Dosovitskiy et al., 2021)[7]に変更することで、精度も向上することも確認できています。

| Model | Backbone | Method | $5^0$ | $5^1$ | $5^2$ | $5^3$ | mean | FB-IoU |
|---|---|---|---|---|---|---|---|---|
| OSLSM | VGG16 | 1-shot | 33.6 | 55.2 | 40.9 | 33.5 | 40.8 | 61.3 |
| co-FCN | | 1-shot | 36.7 | 50.6 | 44.9 | 32.4 | 41.1 | 60.1 |
| AMP-2 | | 1-shot | 41.9 | 50.2 | 46.7 | 34.7 | 43.4 | 61.9 |
| PANet | ResNet50 | 1-shot | 44.0 | 57.5 | 50.8 | 44.0 | 49.1 | |
| PGNet | | 1-shot | 56.0 | 66.9 | 50.6 | 50.4 | 56.0 | 69.9 |
| FWB | ResNet101 | 1-shot | 51.3 | 64.5 | 56.7 | 52.2 | 56.2 | |
| PPNet | | 1-shot | 52.7 | 62.8 | 57.4 | 47.7 | 55.2 | 70.9 |
| DAN | | 1-shot | 54.7 | 68.6 | 57.8 | 51.6 | 58.2 | 71.9 |
| PFENet | | 1-shot | 60.5 | 69.4 | 54.4 | 55.9 | 60.1 | 72.9 |
| RePRI | | 1-shot | 59.6 | 68.6 | **62.2** | 47.2 | 59.4 | |
| HSNet | | 1-shot | **67.3** | **72.3** | 62.0 | **63.1** | **66.2** | **77.6** |
| SPNet | ResNet101 | zero-shot | 23.8 | 17.0 | 14.1 | 18.3 | 18.3 | 44.3 |
| ZS3Net | | zero-shot | 40.8 | 39.4 | 39.3 | 33.6 | 38.3 | 57.7 |
| LSeg | ResNet101 | zero-shot | **52.8** | **53.8** | **44.4** | **38.5** | **47.4** | **64.1** |
| LSeg | ViT-L/16 | zero-shot | **61.3** | **63.6** | **43.1** | **41.0** | **52.3** | **67.0** |

Table 1: Comparison of mIoU and FB-IoU (higher is better) on PASCAL-$5^i$.

| Model | Backbone | Method | $20^0$ | $20^1$ | $20^2$ | $20^3$ | mean | FB-IoU |
|---|---|---|---|---|---|---|---|---|
| PPNet | ResNet50 | 1-shot | 28.1 | 30.8 | 29.5 | 27.7 | 29.0 | - |
| PMM | | 1-shot | 29.3 | 34.8 | 27.1 | 27.3 | 29.6 | - |
| RPMM | | 1-shot | 29.5 | 36.8 | 28.9 | 27.0 | 30.6 | - |
| RePRI | | 1-shot | 32.0 | 38.7 | 32.7 | 33.1 | 34.1 | - |
| FWB | ResNet101 | 1-shot | 17.0 | 18.0 | 21.0 | 28.9 | 21.2 | - |
| DAN | | 1-shot | | | | | 24.4 | 62.3 |
| PFENet | | 1-shot | 36.8 | 41.8 | 38.7 | 36.7 | 38.5 | 63.0 |
| HSNet | | 1-shot | **37.2** | **44.1** | **42.4** | **41.3** | **41.2** | **69.1** |
| ZS3Net | ResNet101 | zero-shot | 18.8 | 20.1 | 24.8 | 20.5 | 21.1 | 55.1 |
| LSeg | ResNet101 | zero-shot | **22.1** | **25.1** | **24.9** | **21.5** | **23.4** | **57.9** |
| LSeg | ViT-L/16 | zero-shot | **28.1** | **27.5** | **30.0** | **23.2** | **27.2** | **59.9** |

Table 2: Comparison of mIoU and FB-IoU (higher is better) on COCO-$20^i$.

## 終わりに

language-driven semantic segmentationというタスクについて紹介させていただきました。現在も画像とテキストを組み合わせた研究は活発に行われており、今後の動向が気になる研究分野かと思います。本記事が最後にとって有益な情報であれば、幸いです。

今後もコニカミノルタAI技術開発部では社会実装まで見据えた技術選定を重視していくことで、価値のあるサービスを提供していけるよう心がけていきます。

## 引用

[1] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Ren´e Ranftl. Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546, 2022.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, pp. 8748-8763, 2021.

[3] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. Advances in Neural Information Processing Systems, 32:468-479, 2019.

[4] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111(1): 98-136, January 2015.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pp. 740-755. Springer, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.

AI技術開発部
人行動、医用画像、外観検査など多様な知的学習モデルの開発に取り組んでいます。

前の記事
FEA它内リッカンカン実験紹介と要望話
次の記事
Japan IT Week 2022 に出展しました

コピペする　ポスト　新ブックマーク 0　Pocket 0　LINEで送る


©2020-2023 Konica Minolta, Inc.