B-02

深層学習による Semantic Segmentation

Deep Learning for Semantic Segmentation

若井大輝†　伊藤滉基†　阪口由宇†

Daiki Wakai　Hiroki Ito　Yu Sakaguchi

1. 抄録

画像をピクセルごとに分類するセマンティックセグメンテーションは、現在、自動運転時の危険認識など先端技術の分野で応用されている。しかし、我々はセマンティックセグメンテーションを高校生の日常生活にも用いられないかと考え、学校授業の欠席者の数を把握するシステムの作成を試みた。画像内の人間を認識するシステムの作成に取り組んだ結果、セマンティックセグメンテーションにおいて十分な精度が得られなかったため、人数を把握することができなかった。以降の研究ではセマンティックセグメンテーションの精度を上げることに努めたい。

2. 背景

深層学習により行う技術の一つであるセマンティックセグメンテーションは、与えられた画像について、それぞれのピクセルを周囲のピクセルの情報に基づいてカテゴリー分類する手法である。自動車の自動運転における危険認識や医療分野にも応用されているが、我々高校生の日常生活にも応用することができないかと考えた。

3. 目的

「欠席把握システム」の作成を目的とする。ただし、ここでの「欠席把握システム」とは、欠席した人が誰であるのかを特定するものではなく、単に人数を把握するものである。また、この際、センサー等を用意しない状態で研究を行うことで、システムの普及を図ると同時に、人数の把握及び個人の特定を行う他の技術との差別化を図った。

4. 方法

学習用のデータセットとして、Cityscapes Dataset[1]の街の画像[一例:図 1]を用いた。左がアーヘン市街地の実際の画像であり、右がラベル付けを行った画像となっている。

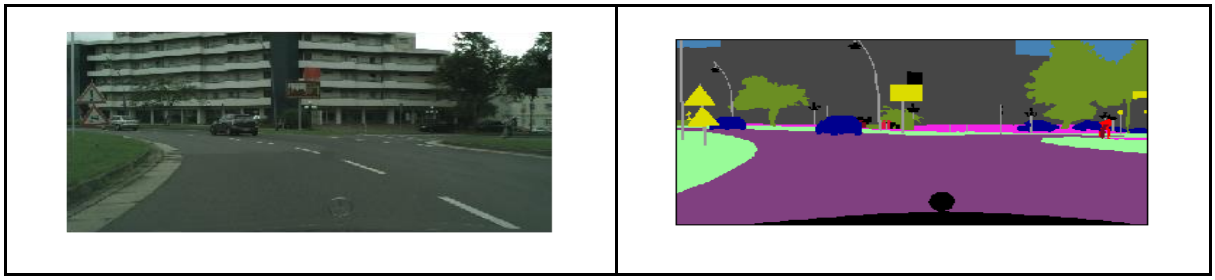

図１　実験に用いた画像の一例。

今回、我々は欠席把握システムの対象として動き回る人々を想定している。そのため、歩いている人々を多く写していると思われる Cityscapes Dataset を用いた。Cityscapes Dataset のうち訓練データ７０９枚、テストデータ５２７枚を用いて、以下のモデル[図２]により学習を行った。このモデルでは言語に python を、ライブラリに keras を採用している。

†西大和学園高等学校　　Nishiyamatogakuen Highschool

また、以下の図では conv は畳み込み層、pool はマックスプーリング層、upsampled　prediction は逆畳みこみ層を示している。

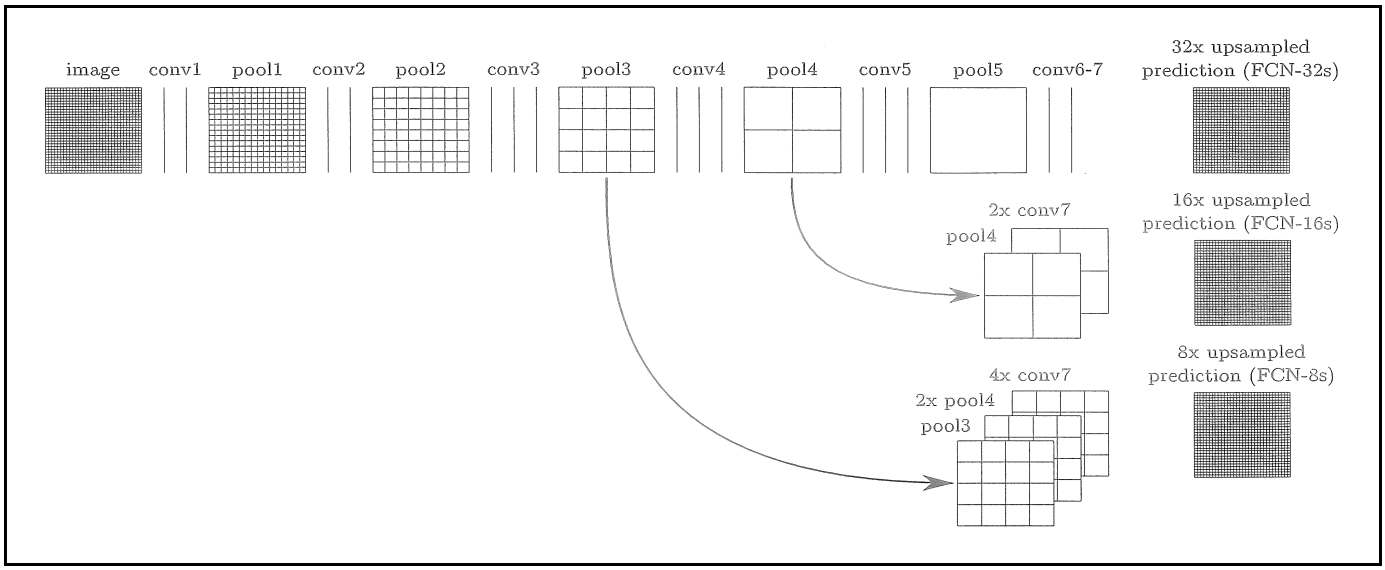

図２　モデルの模式図。

その後、予備実験として重なっている生徒の写真[一例：図３]７枚に対して識別を行えるか調べた後、生徒の教室で撮影した生徒の写真[一例：図４]２９枚における予測を行い、画像内の人間の識別に取り組んだ。

図３　予備実験として使用した画像の一例。

図４　予測に用いた画像の一例。

５．結果

上記のような方法で実験を行ったところ、予測結果[一例：図５]が得られた。以下の図において、左が予備実験の結果の一例として、上記の写真[図３]の結果を、右が教室での実際の画像に対する結果の一例として、上記の写真[図４]の結果を示している。色分けは gist_earth というカラーマップとクラスを対応させており、人間は緑色と黄土色となっている。本来ならば人間を１つのクラスに指定しているため、そのクラスに対応する１色で塗り分けられるはずであるが、原因は未だ不明である。以降の研究では原因の解明と改善に努めたい。

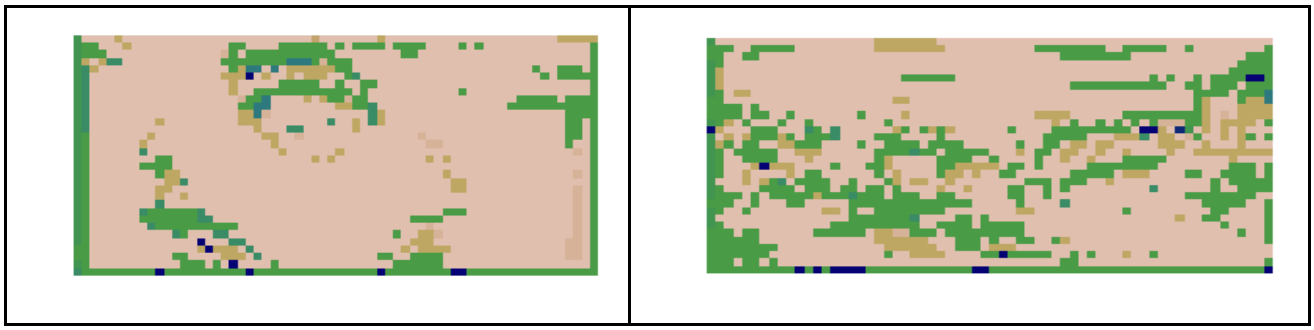

図５　予測結果の一例。

また実験における精度を評価するためのメトリックスとして pixel accuracy を用いた。pixel accuracy は以下の式によって算出される。

pixel accuracy: Σi nii/Σi ti

（ただし nij はクラス i に所属しているが、クラス j に所属していると予測されたピクセルの個数、ti はクラス i に所属するピクセルの総数を指す。）

今回の学習におけるこれらの数値は training のとき 0.4712、 validation のとき 0.1656 であった。これらはあまり高い数値ではないため、欠席把握システムの作成に十分な精度であるとはいえない。

6. 考察

今回の研究では学習段階で、欠席把握システムの作成に対して十分に高い精度を出すことができなかった。これには下記のような原因が考えられる。

１.学習に使った画像の枚数が訓練データ７０９枚、テストデータ５２７枚と少なかったこと

２.学習に使った画像は屋外だが、実際に想定している環境は屋内であること

３.RTX-2080 というマシンを学習に用いたが、マシンの性能が十分な量の学習に適していなかったこと

４.時間的制約からエポックが１００しかなく、十分な回数の試行を行えなかったこと

これらの原因を解決する方法として、教室で生徒を写した写真とラベル付けを行ったものを十分な量用意し、マシンの性能を上げて長時間学習することが考えられる。以降の研究ではこれらの点に留意して、研究を進めていきたい。

7. 謝辞

この度の研究を支援して下さった国際電気通信基礎技術研究所の木戸出教授、このような貴重な機会やアドバイスを頂いた学校の先生方、TA の方に心より御礼を申し上げます。

8. 参考文献

[1]Jonathan Long, Evan Shelhamer, Trevor Darrell Fully Convolutional Networks for Semantic Segmentation in https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf 2019/02/19

[2]M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes Dataset for Semantic Urban Scene Understanding,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016