Vol.2018-CG-171 No.21
2018/9/30

“Pyramid Scene Parsing Network”の実装報告

久山 直輝 [1, a)]

概要：筆者は 2017 年に Hengshuang Zhao らが発表した論文“Pyramid Scene Parsing Network”の実装を行っている．この論文の目的は,様々なシーンでの高精度なセマンティックセグメンテーションを実現することである．提案手法では,様々なデータセットで高精度な結果を得ており,既存手法と比較しても優れていることがわかる．また，この手法は,既存手法に新たな要素を加えた研究となっている．

1. はじめに

本論文[1]では,深層学習を用いたセマンティックセグメンテーションの新しいモデルである PSPNet (Pyramid Scene Parsing Network)を提案している．セマンティックセグメンテーションとは,画像をピクセル毎に分類する技術であり,自動運転や医用画像解析に応用が可能な技術である.また,複雑な物体や様々な物体が存在するシーンの解析は困難であるが,この論文では様々な状況に対応できるネットワークを実現している．

筆者は, [1]の論文を実装することで機械学習の基礎とセマンティックセグメンテーションの基礎を学び, 新たな知見を得ていきたいと考えている．

2. 実装する論文の概要

実装する論文は 2017 年に Hengshuang Zhao らが発表したものである．この論文では, PSPNet (Pyramid Scene Parsing Network)を提案している．PSPNet は大きく分けて既存手法である CNN(Convolutional Neural Network)ベースである部分と論文の最大の特徴である Pyramid Pooling Module の部分を組み合わせた手法である．

ここで CNN について述べる．CNN とは畳み込み層とプーリング層と全結合層から構成されるネットワークである.畳み込み層では局所的な特徴を検出する.プーリング層では特徴を保持したまま画像サイズを小さくすることで情報量を減らしている.全結合層では得られた情報をまとめている.CNN では層が深い程より高度で複雑な特徴を抽出していると考えられている．

本論文では ResNet[2]と Dilated Convolution[3]の既存手法 2 つと,新しい要素である Pyramid Pooling Module を用いている．1 つ目の ResNet[2]は,層がとても深い CNN である.ResNet[2]以前は 20 層程が限度だったが,ResNet[2]は 150 層以上を可能にしているため特徴抽出には良い手法である.2 つ目の Dilated Convolution[3]は,間隔をあけて畳み込みを行うことである.そのためプーリング層の欠点である解像度の低下を防ぐことが可能である.この 2 つの手法を用いる事によってよりよい特徴マップを生成している.そして、Pyramid Pooling Module では,特徴マップを画像の全体と画像の半分と画像の小さい部分,さらに小さい部分の 4 つに分け各大きさで特徴を抽出している．

これらの特徴マップを全て結合することで良い結果をもたらしている.そしてこの論文が優れている事を様々なデータセットを用いて既存手法と比較を行って示している．

3. 実装の進捗状況

現在, ubuntu 上で Python を使用し，実装を行なっている途中である．機械学習のライブラリには, Tensorflow と Keras を用いている．

現時点で，実装環境を整え学習済みのモデルを用いた動作確認ができているがプログラムの内容に関してまだ理解が十分ではない.そのため実行プログラムと訓練プログラムの理解を行っていく予定である.また,この論文を一から訓練学習すると実装にとても時間がかかる.(これは ResNet[2]がとても深い層のため学習に時間がかかるためである.)したがって,今回は Pyramid Pooling Module に焦点をあてて訓練学習を行っていく予定である．そして最終的には自ら訓練したモデルを用いて,実際の出力結果と同等の結果が得られる事を期待している．

参考文献

[1] Hengshuang Zhao,Jianping Shi,Xiaojuan Qi,Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In CVPR,2017.
[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
[3] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122, 2015.

1 静岡大学工学部数理システム工学科
静岡県浜松市中区城北 3-5-1
a) hisayama.naoki.15@shizuoka.ac.jp

ⓒ2018 Information Processing Society of Japan