平成 28 年度卒業論文

再帰型畳み込みニューラルネットワークを用いた風景画像認識の調査

宮崎大学　工学部　情報システム工学科

森　芳雄

指導教員　椋木雅之

目次

1. はじめに

大規模な畳み込みニューラルネットワークによる教師付き深層学習の手法は、一般画像認識などの多くのコンピュータビジョンタスクにおいて非常に成功しており、多くの関心を集めている。一般画像認識問題の 1 つに、画像に写っている対象のクラス名称とその対象が画像のどこに写っているかをピクセルレベルで判別するセメンティック・セグメンテーションがある。このような問題では物体を対象とすることが多く、風景画像の認識を扱っている研究は少ない。物体は同じクラスのものでは形状がある程度は決まっているものが多いが、風景に写っている対象クラスは同じクラスのものでも形が一様でないものもあり、特徴を得にくい。本研究では、セマンティック・セグメンテーションのための手法の一つである、Conditional Random Fields as Recurrent Neural Networks(以下 CRFasRNN と略称)[1]を風景画像に対して適用し、物体に対しては有効であった手法が風景に対しても有効であるかを調査する。

2. 深層学習による一般画像認識の従来手法[2]

ニューラルネットワークは人間の脳の神経回路の仕組みを模したモデルであり、入力層、隠れ層、出力層の 3 つの層に分けられる。入力されたデータは入力層を通り、隠れ層、出力層の順に処理され、出力結果が出来上がる。

深層学習は図 1 のように隠れ層が 2 層以上のニューラルネットワークを対象とした学習手法である。深層学習は情報が深く伝達されるうちに、各層で学習が行われ、この過程で特徴量が自動で計算される。

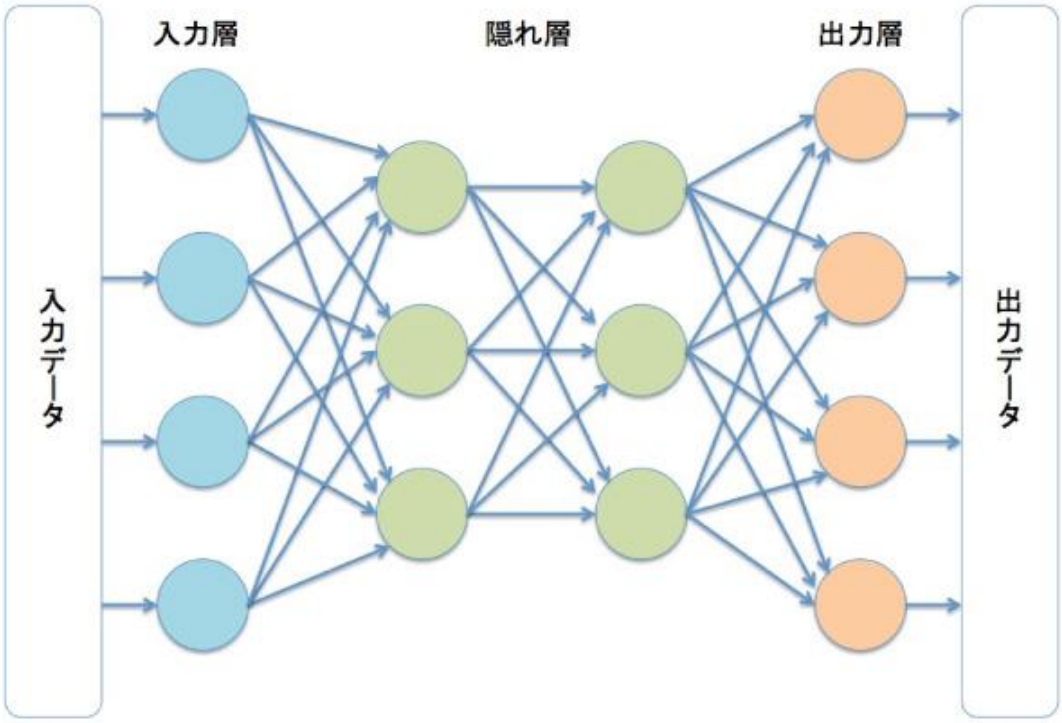


図 1　深層学習のネットワーク構造

http://www.tel.co.jp/museum/magazine/communication/160229_report01_02/03.html

から引用

深層学習を用いた画像認識では、畳み込みニューラルネットワークを使用することが多い。畳み込みニューラルネットワークは、畳み込み層とプーリング層の 2 種類を含む順伝播型ネットワークである。通常のニューラルネットワークは隣接層間のユニットがすべて全結合されているが、畳み込み層とプーリング層では、図 2 のように特定のユニットのみが結合を持つ。

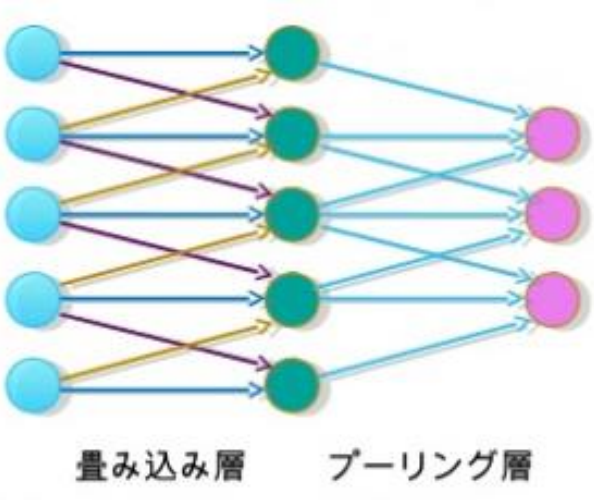


図 2　畳み込み層とプーリング層の結合

https://matome.naver.jp/odai/2140635573608360401/2141120593869192503

から引用

畳み込みニューラルネットワークの典型的な構造は図 3 のようになっている。畳み込み層とプーリング層がペアでこの順に並び、このペアが複数回繰り返され、この後に、正規化層を挿入することがある。畳み込み層とプーリング層の繰り返しの後には、隣接層間のユニットが全結合した全結合層が配置される。畳み込み層では、フィルタと呼ぶサイズの小さい画像の濃淡パターンと類似した濃淡パターンが入力画像上のどこに存在するかを検出することで、フィルタが表す特徴的な濃淡構造を画像から抽出することができる。プーリング層では、畳み込み層で抽出された特徴の位置感度を若干低下させることで、対象とする特徴量の画像内での位置が若干変化した場合でも出力が不変になるようにする。

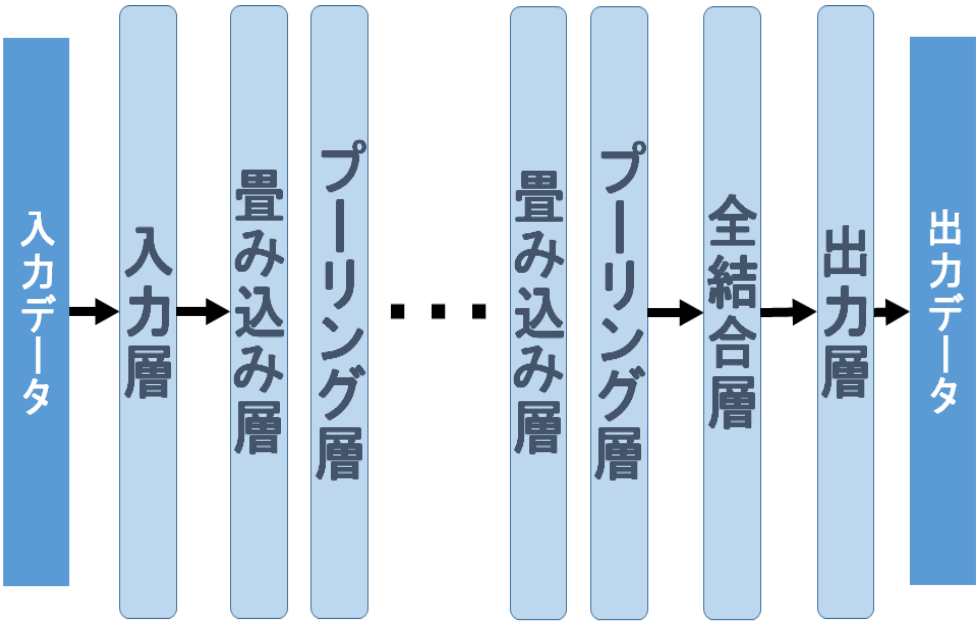


図 3　畳み込みニューラルネットワークの構造

https://thinkit.co.jp/story/2015/09/09/6399 から引用

セマンティック・セグメンテーションに対して、通常の物体認識のために設計された畳み込ニューラルネットワークは、類似ピクセル間のラベル一致、およびラベリング出力の空間的整合性を促進する滑らかさに制約がないため、ピクセルレベルのラベル生成に向いていない。そのため出力が粗くなり、物体の描写が悪くなることがある。畳み込みニューラルネットワークの全結合層の部分を畳み込みに置き換える FCN[3]や畳み込みニューラルネットワークに条件付き確率場を組み合わせた DeepLab[4]は、セマンティック・セグメンテーションで粗い出力が生成される問題に対処しているが、十分ではなかった。

3. CRFasRNN

CRFasRNN[1]はセマンティック・セグメンテーション問題の解決のために提案された手法の一つである。畳み込みニューラルネットワークと条件付き確率場を組み合わせることで粗い出力を改良することができる。CRFasRNN では、条件付き確率場を再帰型ニューラルネットワークとして構築する。

3.1 条件付き確率場

ここでは、条件付き確率場はピクセル間の隣接関係を考えた確率モデルとして考えられている。条件付き確率場は粗いピクセルレベルのラベル予測を精緻化して、鮮明で細かい境界を生成することができ、ピクセルレベルのラベル予測に畳み込みニューラルネットワークを利用する際の欠点を克服することができる。

3.2　再帰型ニューラルネットワーク

再帰型ニューラルネットワークは内部に閉路を持つニューラルネットワークの総称である。再帰型ニューラルネットワークの構造は図 4 のようになっている。再帰型ニューラルネットワークはこの構造のおかげで、情報を一時的に記憶し、振る舞いを動的に変化させることができる。これにより、画像上で隣り合うピクセルの並びをとらえ、分類問題をうまく処理できるようになる。

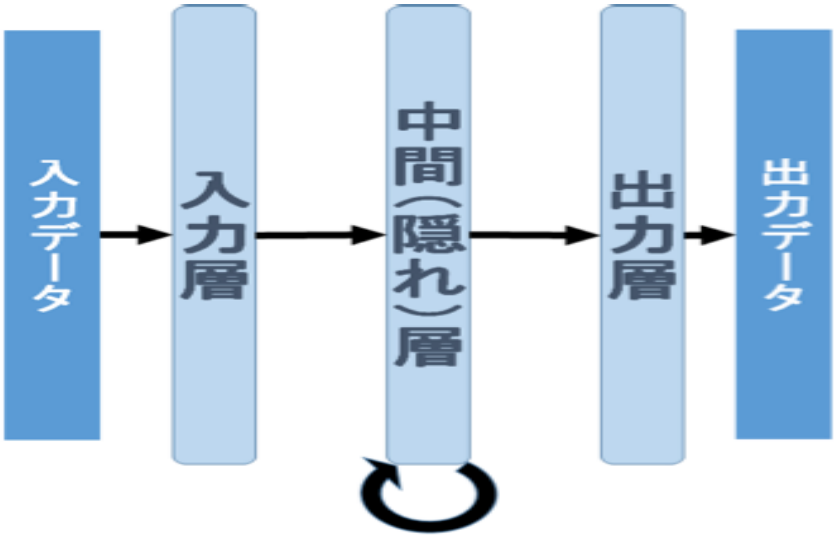


図 4　再帰型ニューラルネットワークの構造

https://thinkit.co.jp/story/2015/09/09/6399 から引用

3.3　CRFasRNN のネットワーク

CRFasRNN は条件付き確率場を再帰型ニューラルネットワークとして構築したものを、畳み込みニューラルネットワークに組み合わせたものである。この

ネットワークは畳み込みニューラルネットワークと条件付き確率場の両方の望ましい特性を有する。

4. 風景画像の認識

風景の認識は物体認識と違い、それぞれの領域の特徴を得ることが難しく、あまり研究されていない。しかし、風景画像も深層学習を用いることで画像に写っているそれぞれの領域を識別できるようになるのではないかと考え、CRFasRNN を用いて風景画像の識別を行う。

4.1 データセット

風景画像の学習と評価を行うための画像は SUN2012 データセット[5]を使用する。このデータセットは、16,873 枚の画像があり、4,919 種類のクラスのものが含まれている。注釈として画像の空や山のようなそれぞれの領域に対して、その領域の頂点の位置と名称が与えられている。

本研究では、この中から空、木、水、山、岩、道路の 6 クラスを識別する。

4.2　画像の作成

CRFasRNN の学習には分類するクラスごとに色分けした画像が必要である。そのため SUN2012 データセットの画像に対して、注釈情報に基づいてクラスごとに色分けした画像を作成する。作成した画像の例を図 5 に示す。

作成したそれぞれのクラスの画像の枚数は、空が 367 枚、木が 45 枚、水が 58 枚、山が 121 枚、岩が 69 枚、道路が 271 枚、合計 931 枚となっている。

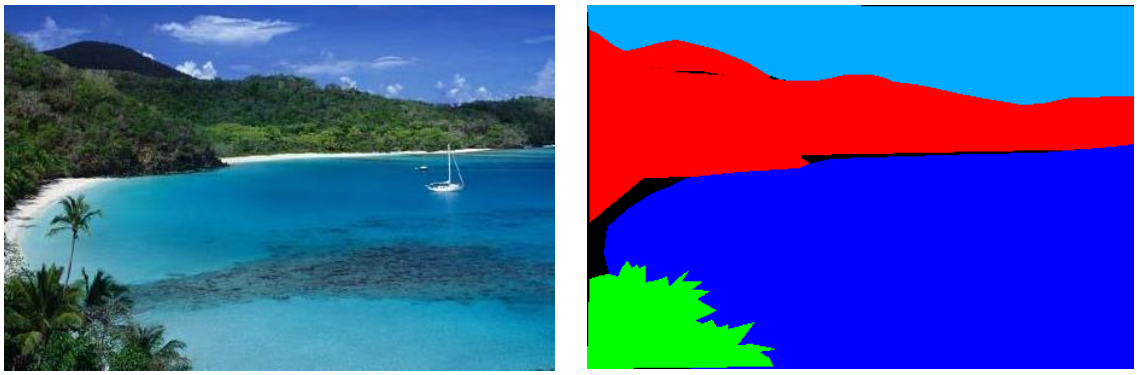


図 5　作成した画像の例

5. 実験

5.1　風景を対象とした実験

風景画像に CRFasRNN を適用し、どれだけ識別できるかを調査する。

5.1.1　実験手順

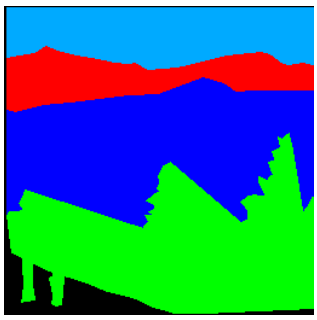
作成した画像とそれに対応した元の風景画像から 841 組の画像を学習用画像として用いて学習を 15,000 回行う。この学習させたネットワークモデルを用いて、入力画像のどこに分類するクラスが写っているか識別した画像を出力する。

識別結果の例を図 6~9 の a)b)e)に、それぞれのクラスを表す色を表 1 に示す。

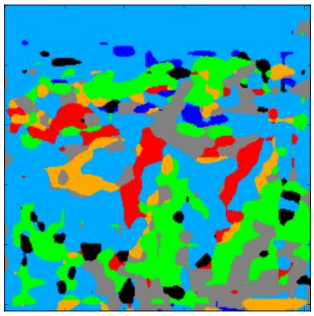
表 1　複雑背景での実験結果

空：	木：	水：
山：	岩：	道路：

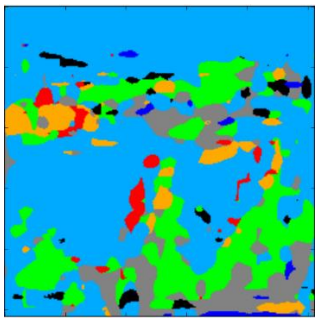
a)入力画像



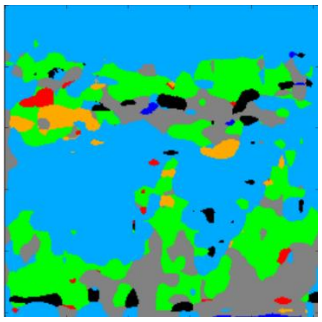
b)正解画像



c)学習 5,000 回



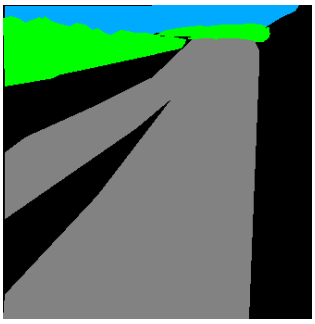
d)学習 10,000 回



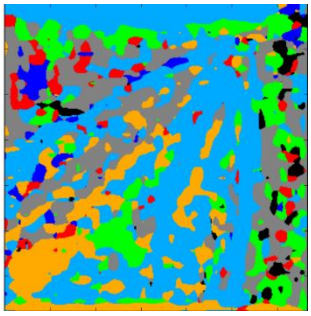
e)学習 15,000 回

図 6　識別結果：例 1

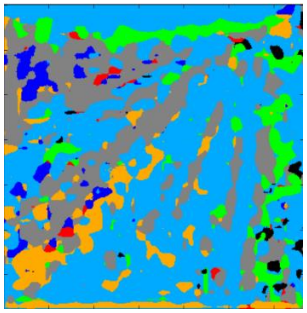
a)入力画像



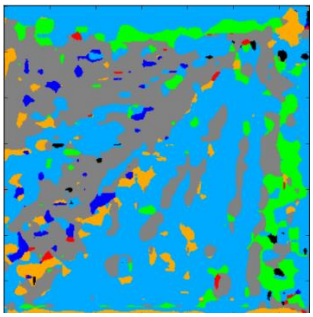
b)正解画像



c)学習 5,000 回



d)学習 10,000 回

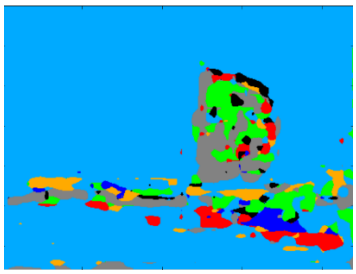


e)学習 15,000 回

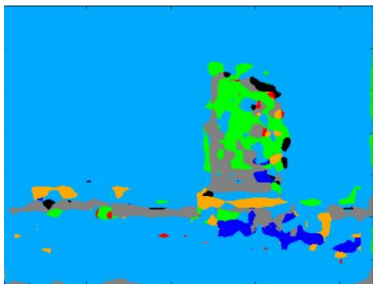
図 7　識別結果：例 2

a)入力画像

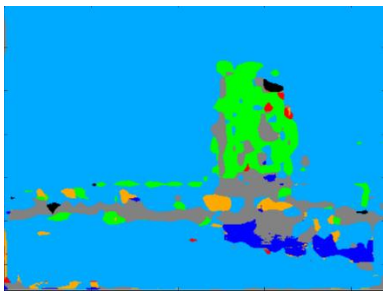
b)正解画像



c)学習 5,000 回

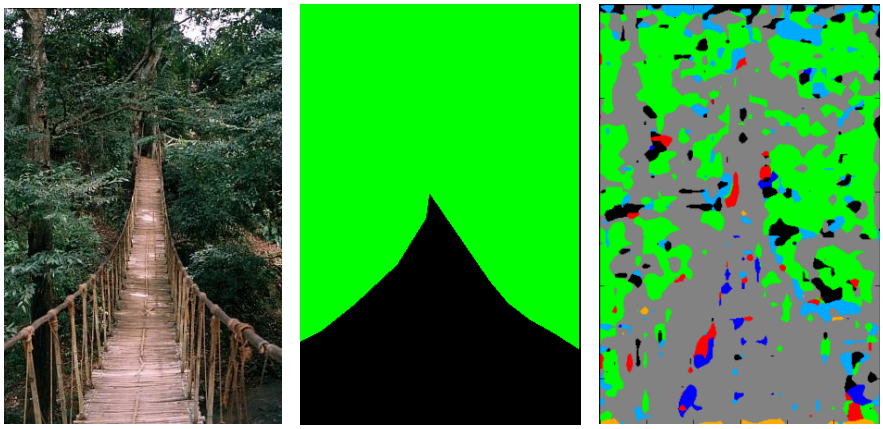


d)学習 10,000 回

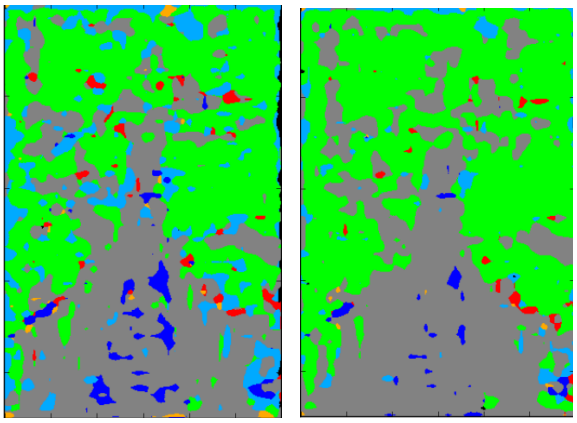


e)学習 15,000 回

図 8 識別結果：例 3



a)入力画像　　b)正解画像　　c)学習 5,000 回



d)学習 10,000 回　　e)学習 15,000 回

図 9 識別結果：例 4

5.1.2 実験結果

図 6、図 7、図 8 の学習 15,000 回では、空と水、道路の領域はその大部分が空として識別されている。ほかのクラスの領域はほとんど識別されておらず、その他として識別されている。図 9 は空のほとんど写っていない画像であり、図 9 の学習 15,000 回では、画像の大部分がその他として識別されていた。学習 15,000 回では、全体的によい結果とはならなかった。

図 10 に学習回数と学習の損失の関係を示す。学習回数が多くなるにつれて小さい値に収束してきている。

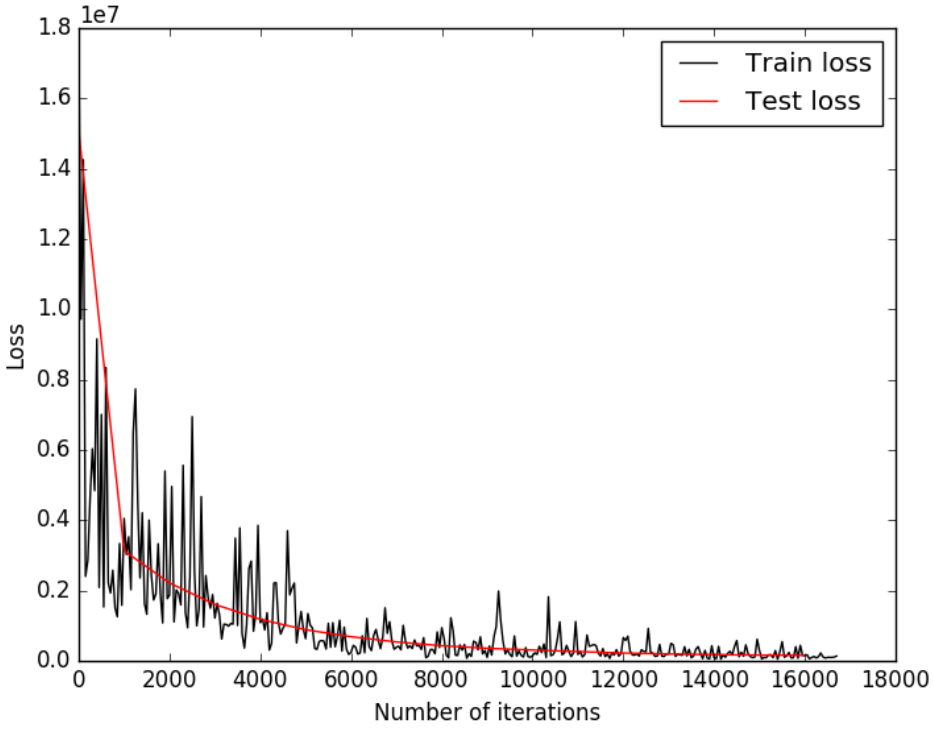


図 10　風景画像の学習の損失

5.2　風景を対象とした学習回数の異なる実験

前節の実験結果より、学習回数とともに損失が小さくなっており、学習が順調に進んでいるように見える。このことを確認するために、学習 5,000 回と 10,000 回のネットワークモデルを用いて実験を行い、学習 15,000 回と比較した。

結果を図 6~9 の c)d)に示す。

図 6 の学習 5,000 回では空と水の領域が空として識別されていた。これは学習 15,000 回と変わらないが、木の領域が木として識別されている点は異なっている。学習 10,000 回では木の領域が他のクラスとして識別されている部分が多くなっている。図 7 の学習 5,000 回では空は正しく識別されている部分が多く、道路も一部は道路として識別されていた。学習 10,000 回になると、道路の大部分が空として識別されるようになった。図 8 の学習 5,000 回、10,000 回では岩の部分にさまざまな識別がされていた。図 9 の学習 5,000 回では、木は正しく識別されている部分が多かったが、学習 10,000 回でさまざまな識別がされていた。

いずれの例でも学習回数が増えるにつれて、識別結果が単調に向上してはいない。

図 10 でも、損失は滑らかに下がってくるわけではなく、ところどころで損失が大きくなる箇所がある。このように損失の収束が一様でなかったため、識別はうまくいかなかったと考えられる。

5.3　物体を対象とした実験

学習の損失の収束が一様でなく、ところどころで損失の大きい箇所が現れるのは、対象が風景であったことが原因か、別の原因があったのかを調査するため、物体を対象とした実験を行った。データセットには文献[1]でも利用している VOC2012[6]を用いる。

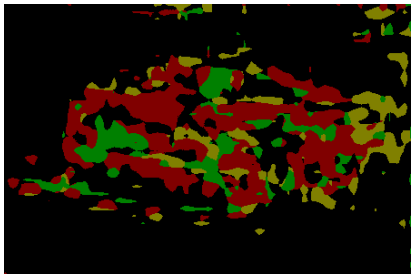
車、バス、バイクの三種類の画像を計 497 枚用いる。このうち 449 枚を学習用画像とした。学習回数は風景の時と同じ、5,000 回、10,000 回、15,000 回で行う。

結果を図 11~13 に示す。また、それぞれの物体を表す色を表 2 に示す。

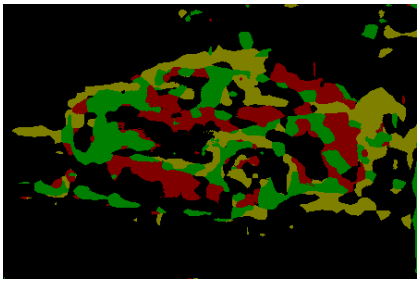
表 2　分類する物体と各物体を表す色

車：	バス：	バイク

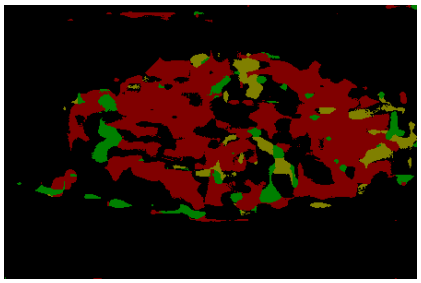
a)入力画像



b)学習 5,000 回



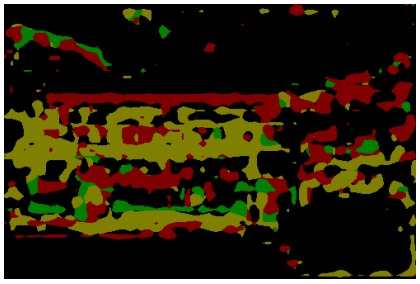
c)学習 10,000 回



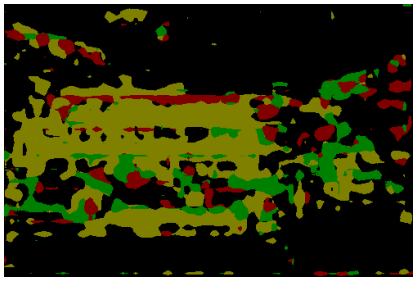
d)学習 15,000 回

図 11　車の画像

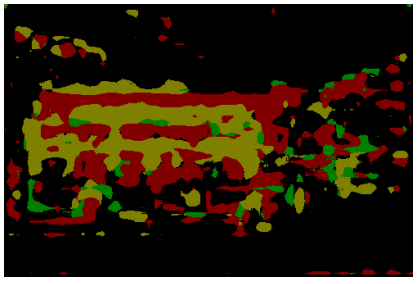
a)入力画像



b)学習 5,000 回



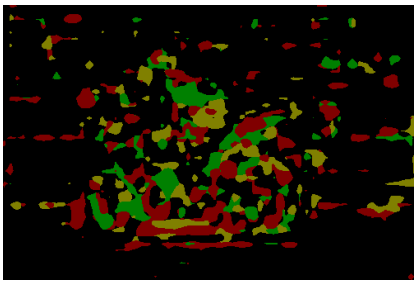
c)学習 10,000 回



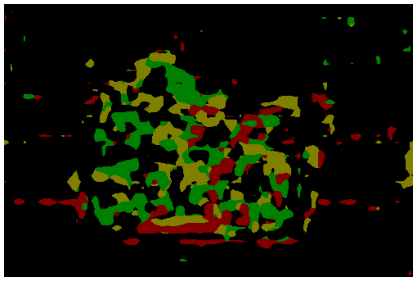
d)学習 15,000 回

図 12　バスの画像

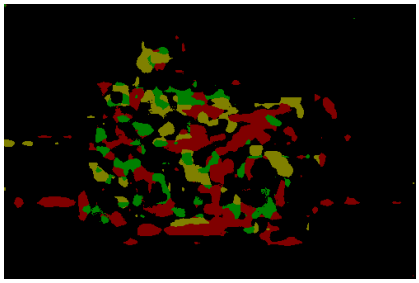
a)入力画像



b)学習 5,000 回



c)学習 10,000 回



d)学習 15,000 回

図 13　バイクの画像

学習 5,000 回では車として識別されている部分が多く、学習 10,000 回ではバスやバイクとして識別されていることが多かった。学習 15,000 回では、多くの画像で車として識別されている部分が多かった。

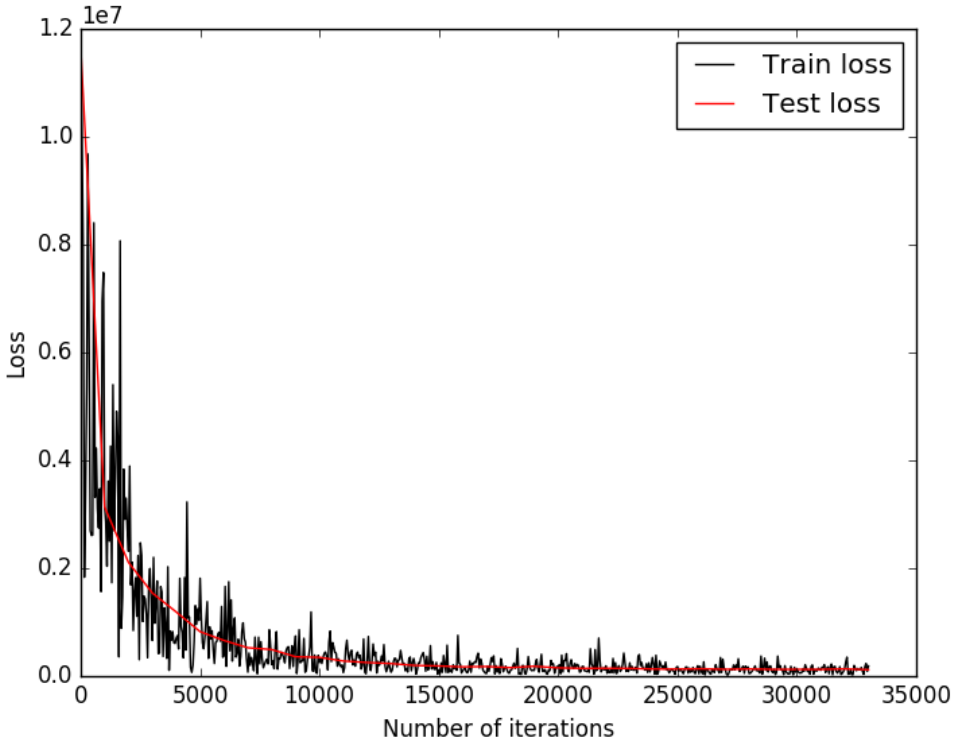


図 14　物体認識の学習の損失

物体識別の学習の損失は図 14 のようになっている。風景画像の認識と同様に物体識別でも損失は一様に収束することはなく、損失が大きくなる箇所があった。このことから風景画像の認識がうまくいかなかったのは対象が風景であるのが原因ではないことが分かった。

6. おわりに

CRFasRNN を用いて風景画像の識別を行った。結果は特定のクラスだけが識別されおり、学習の損失は小さい値に収束していたが、一様に収束していくのではなくところどころ損失が大きくなる箇所があり、よい結果とはならなかった。物体認識の実験も行った結果、風景画像の認識結果が悪かったのは対象が風景であったからではなく、学習回数によって多く識別されるクラスが異なるためであることが分かった。学習の回数が多くなれば、学習の損失はより小さくなっていくと考えられる。今後の課題は、学習回数を多くしていき、風景画像の識別結果がどのように変化していくかを調査することである。

謝辞

最後に、本研究を行うにあたり、お忙しい中ご指導いただいた椋木雅之教授には大変感謝しております。また、椋木研究室の皆様には数々の助言をしていただき、ありがとうございました。本研究では CRFasRNN、Caffe のプログラムを使用させていただきました。両プログラムの製作者の皆様には感謝しております。

参考文献

[1]ShuaiZheng, SadeepJayasumana, Bernardino Romera-Paredes, VibhavVineet, Zhizhong Su, Dalong Du, Chang Huang, Philip H. S. Torr “Conditional Random Fields as Recurrent Neural Networks”, *ICCV*, pp1529-1537, 2015.

[2] 岡谷貴之, “深層学習”, 講談社, 2015.

[3]J. Long, E. Shelhamer, T. Darrell. “Fully convolutional networks for semantic segmentation”, *CVPR*, pp.3431-3440, 2015.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. “Semantic image segmentation with deep convolutional nets and fully connected CRFs”, *ICLR*, 2015.

[5]SUN Database, http://groups.csail.mit.edu/vision/SUN/ (2017/02参照).

[6]VOC2012, http://host.robots.ox.ac.uk/pascal/VOC/voc2012/ (2017/02参照).