人間の視野特性を用いて注視領域と構造情報を考慮した医用画像セグメンテーション

Biomedical Image Segmentation by Retina-like Sequential Attention Mechanism

林 祥平 † Bisser Raytchev † 金田 和文 † 玉木 徹 †
Shohei Hayashi † Bisser Raytchev † Kazufumi Kaneda † Toru Tamaki †
† 広島大学工学研究科ビジュアル情報学研究室

1 概要

本論文では, 医用画像のように扱えるデータが少数の場合に Segmentaiton の精度を向上させるための手法を提案する. 画像全体の中で注視領域を選択的に移動させることによって分類がより困難なクラスの領域部分をより多く処理することを可能とする逐次的な Attention メカニズムを使用する, 生物医学画像の Segmentation のための新しい Deep Learning-based のアルゴリズムを提案する. また各 sub-area 内のクラス情報の空間的分布は、解像度が注目の中心からの距離と共に解像度が向上する網膜のような表現を使用して学習される. 最終的な Segmentation は重複する sub-area のクラス平均をとるアンサンブル学習の効果によって Segmentation の精度を向上させる. 提案手法は Semantic Segmentation タスクにおいて, 従来の patch-based な Convolutional Neural Network(CNN) や医用画像の Segmentation タスクにおいて一般的に利用される U-Net[4] と比較し評価した.

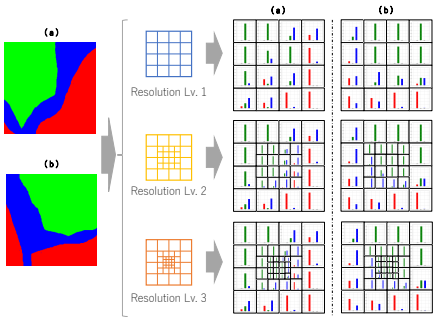

図 1: patch 内に含まれる同じクラス確率の patch を解像度別の小領域に分割した図. この手法では網膜の構造に近い local な sub-area で構造的な情報を学習する. 解像度は中心に近づくにほど高く, 画像内で Attention が移動すると重複する sub-area の情報が組み合わされ, Semantic Segmentation の精度向上が期待できる.

2 はじめに

最近の Deep Learning の手法 [2] において, データ内の複雑な非線形な関係を捉えた階層的な特徴を捉えることによって物体の検出, 認識, Segmentation などの様々な生物医学画像の解析タスクの精度は大きく向上した. 従来の patch-based の手法では予め決められた local patch を画像から取得しネットワークの入力して使用され, その中心にあるピクセルのラベルを教師ラベルとして学習を行い, テスト時には, 学習済みの識別器に patch を入力することで出力層からクラス事後確率を得る. また最近では全結合層を畳み込み層に置き換えた Fully Convolutional Networks (FCN) [5] によって CNN を end-to-end でピクセル単位の学習を効率的に行う手法が patch-based の手法に置き換わり, U-Net などの手法が主流になっている. FCN-based の手法は多くの Segmentation タスクにおいて最先端の精度を示していますが, 高い精度を達成するためには大量のデータセットによって学習する必要があるが, 多くの生物医学画像の Segmentation タスクでは単純にデータが利用出来ない場合や, 専門家によってピクセルレベルで正しいデータを作成するコストが非常に高いため, 扱える画像の量が少ないことが多い. 一方 patch-based の手法では local-patch のみを使用するため, 少ない画像データからでも大量の画像を抽出することが可能である. しかしながら, これらの手法では以下の問題点が存在している.

[i] 従来の patch-based では patch の中心にあるラベルを用いて分類を行なっていたため, patch 内に複数のクラスが含まれる場合, そこに含まれる空間的なクラス構造の情報 (topological imformation) が失われる.

[ii] 少量の画像に対して, 大量の画像を抽出することができるが, 計算量が大きく, FCN-based と比較して精度が低い傾向がある.

[iii] 一方, FCN-based ではデータ量が精度を大きく左右する. また入力された画像の各ピクセルを同等に扱う. これは人間の視覚野の働きと対照的である. 人間は関心のある部分に焦点を当て, その周囲の情報を組み合わせることによってグローバルなシーンを表現を構築するために, 選択的に注視することが知られている [3].

上記の考察に基づき, 本論文では FCN と patch-based の問題点を解決するために, それぞれの手法のメリットを生かした中間的な新しい手法を提案する. 本

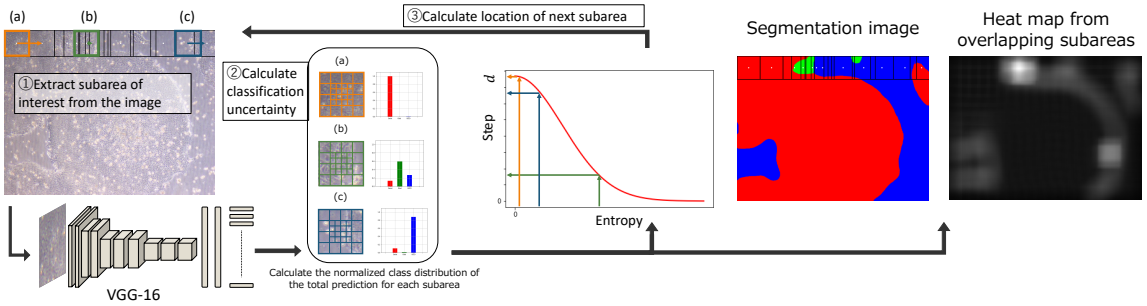

図 2: Attention メカニズムと提案手法の概要図.

手法は patch 内の空間的な構造情報を保持するため, 網膜のように中心窩は高い解像度で patch の中心から離れるにつれて, 解像度は小さくなるように patch を分割する.patch 内のさらに小さな領域を sub-area と呼ぶ. それぞれの sab-area には入力画像の sub-area と対応するようにクラスのヒストグラムを割り当て, patch 内での空間的なクラス構造の情報を学習する. また分類が困難な部分や境界部分がでの予測をより詳細に考慮するように, 注視部分をシフトさせる Attention メカニズムを導入し, 注視すべき領域では小さいステップで移動する. 注視領域では多くの patch が重複するが, 最終的な推定画像は重なり合う領域の各ピクセルのクラス予測を平均することで近隣の全ての重なり合う sub-area からの情報を組み込み (アンサンブル学習), さらに精度が向上する. 以上が本論文で提案する手法の基本的な考え方である. 次節では CNN での実装方法の詳細を述べ, 3 節では提案手法と従来の手法 (center-patch), さらに U-Net による実験結果について示す.

3 提案手法

ここでは, 解像度を用いたクラス分布の手法と Sequential Attention を使った手法について説明する. 解像度レベル ($r = 1$) から $r = 3$ の小領域に分割した例を図 1, Attention メカニズムを用いて注視領域を逐次的に変更しながら推定する手法を 2 に示す. 入力画像から切り出した local patch を, $d \times d \times c$ のテンソル $\mathcal{S}$ とおく. ここで d は patch の一辺の大きさ, c はチャネルを表す (カラー画像なら RGB 値に対応する). 図 1 にあるとおり, patch を解像度別のグリッドで分割し, それぞれの領域でクラス分布のヒストグラム $\boldsymbol{h}^{(i)}$ を計算する. $\boldsymbol{h}^{(i)}$ の k 番目の要素は, i 番目の小領域内 (sub-area) にあるクラス k のピクセルの数を表すものとする.$\boldsymbol{h}^{(i)}$ をそれぞれ合計が 1 になるように正規化して得られるベクトル $\boldsymbol{p}^{(i)}$ を, 小領域 i の確率質量関数 (probability mass function, pmf) として扱う. 次に解像度別に小領域を分割する方法を示す. まず図 1 の上段に示すように $r = 1$ のとき, 4×4 のグリッドに分割する. $r = 1$ のときはグリッドが全て同じ解像度となる. 次にグリッドの中心にある 4 つの sub-area を半分に分割して解像度レベル $r = 2$ を作成し, 解像度が 2 倍の内側の 4×4 の sub-area を形成する. さらに内側の 4 つの sub-area に対して同様のプロセスを続けることで, 高い解像度レベルの sub-area を得る. patch 内の sub-area の数 I は $I = 16 + 12(r - 1)$ で求められる. 初期 sub-area($r = 1$ のとき) は 4×4 である必要はないが, この場合最も中心に近い sub-area は常に 4 であるため, 異なる解像度レベルの作成が特に簡単になる. 本手法ではネットワークの入力として local patch の画像と, 教師データとして対応する sub-area の pmf を学習する. ターゲットの $p^{(i)}$ の pmf と対応する出力ユニット $y^{(i)}$ 間の cross-entropy を損失関数 L として

$$L = -\sum_n \sum_i \sum_k p^{(i)}_{k,n} \log y^{(i)}_{k,n}(S_n; w) \qquad (1)$$

と表せられる. ここで S_n は n 番目に学習に使われれる local-patch のインデックスを表し, i は local-patch に含まれるセルのインデックス, k はクラスを表す. また w はネットワークの重みを表し, 損失関数を最小化することで求められる. 各セルに対応するネットワーク出力ユニットは, 確率を求めるために soft-max 関数を通過する. 最後に本手法で提案する Attention メカニズムについて説明する. このメカニズムは分類が難しい (不確実性が高い) 部分を可能な限り高い解像度かつ, より注意深く推定できるように, 画像全体の中で注視する領域を決定しながら移動をする. 注目している patch S の不確実性を評価するために, 次の関数を使用する.

$$H(S) = -\frac{1}{N} \sum_{i \in S} \sum_k p^{(i)}_k \log p^{(i)}_k \qquad (2)$$

$H(S)$ は予測の不確実性の尺度として, patch の各 sub-area の事後 pmf $p^{(i)}$ から得られた平均エントロピーを表し, 次に注目する焦点までのステップ幅は以下のように求められる.

$$f(H(S)) = -d \exp\{-(H(S))^2 / 2\sigma^2\} \qquad (3)$$

本手法の全体的なプロセスを図 2 に示す. サイズ $d \times d$ ピクセルの patch(図の (a)) を使用して, 入力画像の左上の隅から始める. その patch の不確実性は式 2 で求

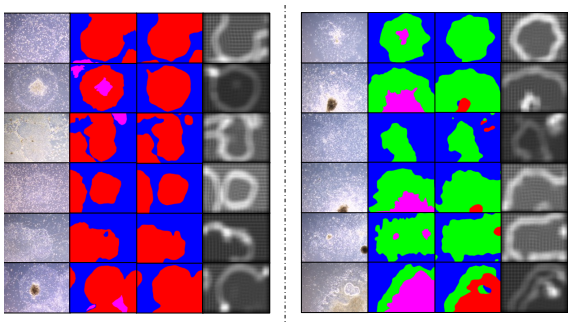

図 3: iPS 細胞のデータセットでの, 提案手法 ($r = 4, d = 192$) の結果.1 列目が元画像, 2 列目が専門家によるラベル付けの真値 (Good は赤, Bad は緑, BGD は青, obscurity は紫で示されている.), 3 列目が提案手法による結果, 4 列目が Attention メカニズムを使用した際の Heatmap.

められ, 移動方向へのピクセル単位のステップ幅は式 3 で求められる. 図 2 の (a) の場合, 中央左のグラフで示すように予測されたクラスの不確実性は 0 となるため, 注視領域は d ピクセル右に移動します. つまりこの例では現在の patch と次の patch に重複はない. 次に (b) の場合には, 事後 pmf の不確実性は非常に高く, 注視領域はわずかに右に移動することでこの領域の周囲を高い解像度で評価できる. 図 2 の Heatmap で示されるように, 不確実性が高いクラスの境界部分に多く重複することを示している (重複度に比例して白くなる). このプロセスは画像の右端に達するまで繰り返され, 右端に達すると, 垂直方向に 10 ピクセル移動した左端に移動して次の行をスキャンし, 画像全体が処理されるまで繰り返される. 画像全体で水平方向のスキャンを終えると, 垂直方向も同様のプロセスを行う. この Attention メカニズムは画像全体で注視領域を移動するが, 各 patch に対応する sub-area の事後クラス pmf は画像と同じサイズの確率マップに保存される. 画像内の各ピクセルには、そのピクセルの上に配置されたグリッドのセルの pmf が割り当てられる. 複数の patch が重複する部分では, ピクセルごとに部分的に重複するすべての sub-area の pmf をピクセルごとに平均化することにより, 確率マップが計算される. 最終的にピクセルのクラスは図 2 の Segmentation Image で示されるように, 確率マップから最も高い確率を持つクラスが割り当てられる.

4 実験

ここでは構造情報を用いた提案手法と, 従来の分類による手法 (patch-center) を比較し, 評価した. また同時に，Semantic Segmentation 用の深層モデルのベースラインとして用いられる U-Net についても比較のために実験を行った. 全てのモデルで事前学習は行わず, スクラッチから学習を行った. 通常 U-Net などの深層モデルは画像全体を入力として用いるが, それに加えて構造情報を表現する一つの方法として, patch 単位で U-Net を学習する実験も行った. それぞれの手法をここでは UNet-image と UNet-patch と呼ぶ. 実験では, 位相差顕微鏡を搭載したデジタルカメラで撮影した iPS 細胞 [7] のコロニー画像 59 枚を使用した. 各画像は専門家によって領域ごとに Good(未分化の細胞), Bad(分化した細胞), Background(BGD, 培養液) と, 専門家から見てもわからない部分 (obscutiry) の 4 つのクラスにラベル付けされている. 画像は全て 1600×1200 ピクセルである. 学習・推定は Good, Bad, BGD の 3 クラスに対して行い, 評価時は obscurity の部分は除外した. 図 3 でデータセットの画像と専門家による真値画像の例を示す. ネットワークは VGG-16[6] をベースとして変更を加えた CNN を用いた.VGG-16 では畳み込み層が 13 層あるが, ここでは 10 層の畳み込み層を含むネットワークを使用した. また, VGG-16 では最後の全結合層のユニット数は 4096 となっているが, 提案手法では 1024 ユニットの全結合層を用いた. また提案手法の出力層のユニットは $N \times k$ 個のユニットを使用し, 単一のクラスラベルを推定するのではなく, patch 内のクラスの分布を解像度に合わせて推定する. 学習率は 0.0001 で固定した.batch size は 16 とし, 20 エポック学習させた. patch のサイズは $d = 128, d = 192, d = 256$ の 4 つで実験を行い, 学習時の patch のステップ幅はいずれも 45 ピクセルで固定した. また提案手法の解像度レベルは $r = 1$ から $r = 5$ で実験した. 分類での手法では, 中心ピクセルのラベルをその patch のラベルとして学習した. そのラベルが obscurity であった patch は学習には使用しなかった.

UNet-image は学習率 0.0001, batch size は 1 で 100 エポック学習した. ただしアスペクト比を保ち, 短辺部分はパディングを施して, 1024×1024 ピクセルにリサイズした画像を用いて, 学習・推定した. 実装には Keras[1] を使用し, GeForce GTX1080 Ti を使用した. 各手法で得られた推定結果は, Jaccard Index, Dice coefficient, True Positive Rate(TPR), True Negative Rate(TNR), Pixel Accuracy で評価した. 5-fold cross validation によって評価を行い, 得られた結果を表 1 に示す. 3 に推定結果の例を示す. 一列目は入力画像, 二列目がそれに対する真値, 三列目と四列目が提案手法と Attntion メカニズムによる Heatmap よる推定結果である. Heatmap から分かるように, 境界部分や複数のクラスが混在する部分で注視していることが分かる. また式 3 で与える σ を変化させた場合の精度と 1 枚の画像の推定に必要な patch の枚数を図 4 に示す. Attention メカニズムを使用せず step を 10 ピクセルで固定した場合, $d = 192$ のとき, 全体画像 1 枚あたり 25200 枚の patch が必要であるが, Attention メカニズムを使用した場合, 適切な σ の値を選んだ場合, 少ない枚数にも関わらず高い精度で認識できることが分かる.

表 1: iPS dataset での結果.

Method	Patch size	Jaccard Index	Dice	TPR	TNR	Accuracy
patch-center	$d = 128$	0.811 ± 0.029	0.878 ± 0.026	0.873 ± 0.028	0.917 ± 0.012	0.931 ± 0.012
ResLv-1		0.812 ± 0.038	0.878 ± 0.032	0.874 ± 0.033	0.918 ± 0.020	0.935 ± 0.012
ResLv-2		0.815 ± 0.039	0.881 ± 0.032	0.878 ± 0.034	0.923 ± 0.015	0.934 ± 0.014
ResLv-3		0.816 ± 0.030	0.881 ± 0.029	0.879 ± 0.033	0.920 ± 0.011	0.935 ± 0.008
ResLv-4		0.818 ± 0.034	0.882 ± 0.031	0.881 ± 0.031	0.921 ± 0.017	0.936 ± 0.011
ResLv-5		0.826 ± 0.032	0.891 ± 0.030	0.890 ± 0.032	0.924 ± 0.018	0.936 ± 0.009
UNet-patch		0.810 ± 0.035	0.879 ± 0.030	0.873 ± 0.034	0.921 ± 0.014	0.933 ± 0.012
patch-center	$d = 192$	0.778 ± 0.029	0.856 ± 0.025	0.849 ± 0.020	0.899 ± 0.017	0.917 ± 0.017
ResLv-1		0.810 ± 0.037	0.876 ± 0.030	0.872 ± 0.031	0.914 ± 0.021	0.933 ± 0.016
ResLv-2		0.817 ± 0.041	0.880 ± 0.038	0.879 ± 0.041	0.922 ± 0.017	0.936 ± 0.011
ResLv-3		0.821 ± 0.032	0.885 ± 0.030	0.883 ± 0.032	0.926 ± 0.010	0.935 ± 0.011
ResLv-4		**0.831 ± 0.036**	**0.894 ± 0.030**	**0.890 ± 0.034**	**0.926 ± 0.015**	**0.940 ± 0.012**
ResLv-5		0.825 ± 0.032	0.887 ± 0.030	0.883 ± 0.032	0.925 ± 0.015	0.938 ± 0.012
UNet-patch		0.809 ± 0.036	0.878 ± 0.029	0.870 ± 0.034	0.920 ± 0.015	0.933 ± 0.015
patch-center	$d = 256$	0.732 ± 0.038	0.822 ± 0.037	0.813 ± 0.039	0.862 ± 0.023	0.897 ± 0.019
ResLv-1		0.804 ± 0.039	0.871 ± 0.035	0.866 ± 0.036	0.910 ± 0.018	0.931 ± 0.012
ResLv-2		0.810 ± 0.038	0.877 ± 0.037	0.870 ± 0.040	0.918 ± 0.018	0.932 ± 0.012
ResLv-3		0.819 ± 0.029	0.882 ± 0.028	0.879 ± 0.034	0.922 ± 0.016	0.937 ± 0.010
ResLv-4		0.814 ± 0.033	0.877 ± 0.030	0.871 ± 0.031	0.917 ± 0.020	0.936 ± 0.011
ResLv-5		0.811 ± 0.034	0.879 ± 0.028	0.874 ± 0.030	0.921 ± 0.014	0.933 ± 0.012
UNet-patch		0.791 ± 0.050	0.864 ± 0.039	0.853 ± 0.050	0.920 ± 0.013	0.924 ± 0.022
UNet-image		0.806 ± 0.033	0.877 ± 0.029	0.874 ± 0.031	0.919 ± 0.013	0.928 ± 0.008

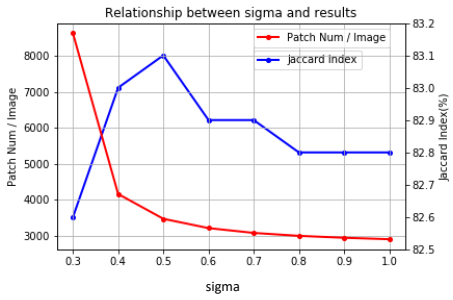

図 4: sigma の値と Jaccard Index, 1 枚の画像を推定する際に必要な patch の枚数

5 結論

本研究では, local patch に対して単一のクラスラベルを与える代わりに, 小領域ごとに計算された確率質量関数を教師データとして与える CNN モデルを提案した. これにより従来手法では使われていなかった, patch 内に含まれるクラスの構造的な情報を学習可能になった. また Attention メカニズムを導入し, 推定に必要な patch の枚数を減らすことで効率の良い推定を行い, さらに境界部分や複数のクラスが混在する領域を注視することで高い認識精度で推定することが可能となった. 今後の課題として, 他のデータセットに対しての検証や, 既存のモデルとの性能比較が挙げられる.

参考文献

[1] F. Chollet. *keras*, 2015. Available at:`https://github.com/fchollet/keras`.

[2] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[3] R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1-3):17–42, 2000.

[4] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[5] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017.

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[7] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.