東京大学
情報理工学系研究科 創造情報学専攻
修士論文

病理画像解析支援のための深層能動学習に関する研究

Cost effective learning of pathological images with deep active learning

福田　圭佑
Keisuke Fukuta

指導教員　原田 達也 教授

2018 年 1 月

概要

病理標本を撮像して診断を行うデジタルパソロジー技術の普及に伴い，画像認識技術によって病理医の診断を補助するための研究が数多く行われている．一方，高精度な識別率を達成する学習モデルを獲得するには，大規模なラベル付きデータを構築する必要がある．病理画像等の医療画像の学習データセットの作成には高度に専門的な知識が不可欠であるため，医師にとって大きな負担となってしまいアノテーションコストが非常に高い．そこで本研究では，画像群の中からモデルの識別精度向上に寄与する可能性の高いサンプルのみにアノテーションを付与することでアノテーションコストを抑えつつ高精度なモデルを獲得する能動学習の枠組みを採用する．また，能動学習を病理画像に適用するための工夫だけでなく，近年の画像認識分野で目覚ましい成果を挙げている Convolutional Neural Network (CNN) を能動学習に適用するための新たな手法を提案し，大規模病理画像データセットに対して実験を行い有効性を検証した．

Abstract

This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract.

This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract.

This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract.

This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract.

This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract. This sentence is for English abstract.

目次

第 1 章　序論 1
1.1　研究の背景 1
1.2　研究の目的 2
1.3　研究の構成 2
1.4　研究の貢献 2

第 2 章　関連研究 3
2.1　病理画像解析 3
2.1.1　テクスチャとしての性質 3
2.2　深層能動学習 3

第 3 章　能動学習 4
3.1　基本的な概念 4
3.2　能動学習の適用される状況 5
3.3　選考基準 5
3.3.1　Uncertainty Sampling 5
3.3.2　Extracting Data Structure 5
3.3.3　Query By Committee 5
3.3.4　Expected Model Change / Expected Varance Reduction 5

第 4 章　本論 6
4.1　次 6
4.1.1　下 6
4.1.2　その次 6
4.1.3　そのまた次 7
4.2　最後 7

第 5 章　結論 8

参考文献 9

付録 A ソースコード 13

第 1 章

序論

1.1 研究の背景

(デジタルパソロジー 診断補助の可能性)

病理標本のデジタル画像を撮影し，ディスプレイに表示することで細胞の組織を観察して診断を行うデジタルパソロジー (Digital Pathology) の普及が進んでいる．病理画像は 1 スライド当たりギガピクセルに及ぶ非常に高解像画像で，組織すべてを隅々まで確認して異常がないか診断を行うのは熟練の病理医でも時間を要する．一方，デジタルパソロジーの普及に伴い組織の画像データが大量に保管されるようになり，画像解析技術の応用によって病理医の診断を補助するための研究が数多く行われている．[1]，ToDo: 引用

(機械学習による診断, 特に CNN, これらはめちゃくちゃデータ食う)

[2], [3]

機械学習手法を用い，汎化性能の高い識別器を教師付き学習によって構築するには大量のラベル付きデータが必要となることが多い．また，近年の画像認識分野において，Convolutional Neural Network (CNN) と呼ばれる多層ニューラルネットワークを用いた機械学習技術による成果が目覚ましい．従来の画像認識では人手によって特徴量を設計し，その特徴量を用いて識別器を学習していたのに対し，CNN は学習の過程で訓練データから識別に有効な特徴量を抽出する表現学習と識別器の学習を同時に行うことができる点が特徴的である．

病理画像認識の分野においても CNN を使用した研究は急速に増加しており，組織の異常や病変の認識，検出の精度は熟練の病理医に匹敵することが示されている．1 スライド当たりギガピクセルに及ぶ病理画像解析では，スライドを複数の画像パッチに分割し，それらに対して識別及びセグメンテーションを行う場合が多い．我々は，Camelyon17 という乳癌のリンパ節転移を検出するタスクにおいて，CNN の共分散特徴量を利用した学習器によって，病理医に近い精度を達成した．また，機械による識別結果を利用しながら病理医が診断を行った場合，機械のみ，病理医のみによる診断と比較して速度，精度ともに向上したという報告がある．

(医療画像のようにアノテーションコスト高いと厳しい (ここで camelyon)) これらのことから，CNN を病理画像に利用するのは非常に有望であると考えられるが，これらの学習には大量のラベル，またはアノテーションを必要であるという大きな欠点がある．当然のことながら

病理画像の学習データの作成には高度に専門的な知識を必要とするため，ギガピクセルに及ぶ画像の詳細なアノテーションの付与には莫大なコストが伴う．

(アノテーションコストが高い問題を緩和するための手法の一つに active learning がある)

ラベルが不足した状況に対する機械学習のアプローチとして，半教師付き学習や教師なし学習などが挙げられるが，教師あり学習と比較して精度の点で大きく劣る場合が多い．

教師付き学習は，入力（質問）と出力（答え）の組からなる訓練データを用いて， その背後に潜んでいる入出力関係（関数）を学習する問題である教師なし学習は，文字通り教師がいない状況での学習であり， 出力（答え）の無い入力データのみが与えられる． 教師なし学習の目的は状況によって異なり，数学的にきちんと定式化できない場合が多い． 例えば，入力データの似たもの同士をグループ化するクラスタリングがその典型的な例である教師付き学習では入力と出力の組からなる訓練データが与えられ， 教師なし学習では入力だけの訓練データが与えられる．半教師付き学習は，これらの中間の状況に対応する学習問題であり， 入力と出力の組からなる訓練データに加え入力だけの訓練データも与えられる．半教師付き学習の目標は，教師付き学習と同じく高い汎化能力を獲得することである．半教師付き学習では， 入出力両方が揃っている訓練データの数は少なく， 入力だけの訓練データの数は非常に多い場合を考えるのが典型的である．このような状況では，少数の入出力データだけでなく多数の入力データも用いる事により， より高い汎化能力が獲得できると期待される．

能動学習は

そこで本研究では，病理医が支払うコストを最小にしつつ精度を担保するためのアプローチとして，学習データのラベルを学習の過程で機械とアノテーターとの interaction を通じて増加させていく Active Learning を病理画像解析に取り入れることを提案する．

(active の説明 理論的にもいろいろあるぜ)

(active + Pathology) sample, i.i.d じゃないよね．似た領域には似たサンプルがあるよね enforce diversity

(active + deep)

このような学習方法が浸透するにつれ，大規模データベースに対するアノテーションコストは増加している．

1.2　研究の目的

1.3　研究の構成

1.4　研究の貢献

第 2 章

関連研究

2.1 病理画像解析

2.1.1 テクスチャとしての性質

2.2 深層能動学習

第 3 章

能動学習

3.1 基本的な概念

ラベルなしデータ集合 $\mathcal{U} = \{x_1, \cdots, x_n\}$ を与えられた時に

能動学習，Active Learning[4] とは，機械学習における一つの枠組みである．一般に，教師つき学習において識別精度の高いモデルを学習させるためには膨大なアノテーション付きデータが必要となる．そこで，能動学習では，得られた大量のラベルなしデータから最もモデル更新に寄与する可能性のあるサンプルを調べ，それにアノテーションを付与することで，少ないアノテーションコストのもとでいかに高精度な学習モデルを作成するかを目的としている．

機械学習では，教師データ作成コストが問題となる場合が多く，少量のデータから効率的に学習する枠組みが重要視されている．少量のデータからの学習として，能動学習 (Active Learning)，転移学習（transfer learning) が提案されている．能動学習は，教師データの作成コストが大きい場合に用いられる枠組みで，一部のデータのみに選択的にラベルを付ける方法である．

学習に用いられる識別器としては，データ x が与えられら時のラベルが正例である確率が定義されているもの．

能動学習の基本的な流れを以下に示す

1. モデル更新に寄与する可能性のあるサンプルを選択
2. 人手，専門家によってアノテーションを付与
3. 付与されたアノテーションを利用し教師あり学習

学習が飽和するまでこれらのサイクルを繰り返す

学習の精度はデータ量の対数スケールに比例して上昇精度向上に要求されるデータ量は指数的に増大

教師あり学習：問題集と解答集半教師付き学習：教科書と章末問題能動学習：先生にわからないところを聞く

その選考基準は様々に考案されている．

サンプリングバイアス能動学習で得られたデータ集合は実際のデータ集合とは異なる

3.2 能動学習の適用される状況

Membership Query Synthesis

Membership Query Synthesis はストリーミングデータ (サンプルが次々に入力されていくような場合) に対するアプローチで，ストリーミングデータを直接，人間に提示してラベルを付けるのではなく，1 つまたは複数のサンプルから新しいサンプルを生成し人間に提示する事でラベル付けを行う方法である．図 1.3 に Membership Query Synthesis の流れを示す．

Stream-Based Selective Sampling

Stream-Based Selective Sampling は Membership Query Synthesis と同じストリーミングデータに対する手法であるが，入力されるストリーミングデータに対してそれぞれのサンプルにラベルを付けるかを判断し，ラベルを付けると判断しサンプルを人間に提示する方式である．図 1.4 に Stream-Based Selective Sampling の流れを示す．

Pool-Based Sampling

上記の 2 つがストリーミングデータに対するものであったが，Pool-Based Sampling はまとまったサンプルに対するアプローチである．サンプルプール (複数のサンプル) を扱うためサンプルの持つ情報量の計算が容易でである．また，入力サンプルの確率分布を予測することもでき，プールの中からサンプルを選択する方法である．ただし，サンプルの保存や，大きなサンプルプールに対して計算を行う必要があるため機器のストレージやメモリ，演算能力などが求められる．このためモバイル機器のような機器では扱えない場合がある．図 1.5 に Pool-Based Sampling の流れを示す．

3.3 選考基準

3.3.1 Uncertainty Sampling

3.3.2 Extracting Data Structure

3.3.3 Query By Committee

3.3.4 Expected Model Change / Expected Varance Reduction

第 4 章

本論

詳細は，表 4.1 を参照.

4.1 次

4.1.1 下

4.1.2 その次

これは，本論の文章である．これは，本論の文章[*1]である．これは，本論の文章である．これは，本論[*2]の文章である.

$$\sum_{k=1}^{n} = \frac{n(n+1)}{2} \tag{4.1}$$

式 (4.1) より，結論が得られる．詳細は，図 4.1 を参照.

表 4.1. 表のタイトル

列 1	列 2	列 3
項目 a1	項目 a2	項目 a3
項目 b1	項目 b2	項目 b3
項目 c1	項目 c2	項目 c3

[*1] 脚注はこのように書く.

[*2] 脚注を入れすぎると読みにくくなるという意見もある．長文の脚注も避けるべきであるとの主張もある．適切な脚注になっているかどうか，十分検討すべきである.

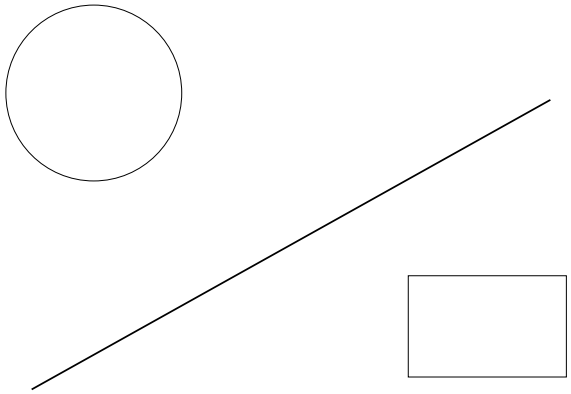

図 4.1. 図のタイトル

4.1.3 そのまた次

4.2 最後

第 5 章

結論

これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．

これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．

これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．

これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．

これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．これは，結論の文章である．

参考文献

[1] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, Vol. 2, pp. 147–171, 2009.

[2] Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pp. 496–499. IEEE, 2008.

[3] M Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N Gurcan. Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering*, Vol. 58, No. 7, pp. 1977–1984, 2011.

[4] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, Vol. 52, No. 55-66, p. 11, 2010.

謝辞

TODO

付録 A

ソースコード

```
int main () {
  ...
  ...
}
```