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 本研究の背景

(デジタルパソロジー 診断補助の可能性) 病理診断は、病理医がスライドガラス上の標本を観察することにより行われる。近年では標本全体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の普及が進んでいる [?]. このデジタル画像は Whole slide image, WSI と呼ばれる。WSI 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.

(機械学習による診断) 一方，デジタルパソロジーの普及に伴って WSI が大量に保管されるようになり，機械学習を利用した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 [?]，[?]、[?]。中でも、WSI を入力として対応する病名を出力する自動診断に関する研究が盛んに行われている [?]、 [?]。

(特に CNN, これらはめちゃくちゃデータ食う) しかし、近年の画像認識分野において目覚ましい成果を挙げている Convolutional Neural Network (CNN) をはじめとするように、機械学習手法を用い，汎化性能の高い識別器を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.

(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)) 病理画像など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，医師への大きな負担となり、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．また、特に病理画像においては、1 枚ギガピクセルに及ぶ WSI に対する詳細なアノテーションの付与には莫大なコストが伴う．Camelyon Grand Challenge[?] にて公開された Camelyon Dataset では、1000 枚にも及ぶ WSI に詳細な乳癌のリンパ節転移が生じている領域にアノテーションが付与されており、WSI1 枚当たり熟練の病理医が 1 時間を費やしたという。

(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) 教師データ作成コストが問題となるのは医療画像解析に限った話ではなく、機械学習では少量のデータから効率的に学習する枠組みの研究も盛んに行われている。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，能動学習と呼ばれる枠組みがある。

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り、しばしば問題集と解答集が与えられてテスト問題でできるだけ高い点を取る状況と例えられる。また、半教師付き学習は問題集と一部の解答のみが与えられる状況だと言える。ここで、能動学習とは、教師なし学習と同様に問題集のみが与えられるが、それとは別に教師に自分のわからないところを聞くことができる枠組みであると例えることが出来る。能動学習は，「もし学習アルゴリズムが学習データ全体の中から任意のデータを選択することができる場合，適切な選択によって得られる学習器の性能は向上する」という仮説に基づいている。すなわち、大量に与えられたラベルなしデータから中から、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を選択し、オラクル (アノテーター、もしくはドメインの専門家) のみにアノテーションをその都度付与してもらうことで、モデルの識別率を犠牲にすることなくアノテーションコストを最小化するためのアプローチである。サンプル選択における理論的な研究だけでなく、実用的なアプリケーションに関する研究も近年増加しており、高いアノテーションコストが大きな課題である病理画像解析への適用親和性は非常に高いと言える。

1.2　本研究の目的

本研究では，病理画像解析において病理医が支払うコストを最小にしつつモデルの識別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能動学習を採用したシステムの構築を目的とする。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，その有効性を検証する。

1.3　本研究の構成

本論文の構成を以下に示す.

第 1 章で本研究の背景と目的を述べた。

第 2 章では，本研究と関わりのある病理画像解析、深層能動学習の先行研究について述べる。

第 3 章では，能動学習の基本的な概念を述べる。

第 4 章では、本研究で提案する手法について述べる。

第 5 章では，実験の目的，設定，結果について述べる.

第 6 章では，本研究の結論，および展望について述べる.

1.4　本研究の貢献

- 病理画像解析に能動学習を利用するための実用的な手法の提案

- 深層能動学習のための有効なサンプル選択の戦略の提案

第 2 章

関連研究

2.1 病理画像解析

本節では、病理画像解析特有の性質と関連研究を述べる。

2.1.1 概観

第 1 章で述べたように、病理画像とは WSI と呼ばれる巨大なデジタル画像のことを指す。ToDo: 画像病理画像解析のおけるアプリケーションは、大きく分けて 3 つに分類される [?]。自動診断による医師の補助、類似画像検索による医師の補助、画像と画像以外の情報との関連性の解析、が挙げられる。近年の画像認識技術の向上による、自動診断に関する研究が急速に増加している。本研究でも、自動診断に位置づけられる癌の自動検知タスクに焦点を当てる。

2.1.2 パッチベースの分類

一般に、画像認識タスクにおいて用いられる画像はせいぜい 1000 × 1000 以下である。しかし、数万ピクセル×数万ピクセルに及ぶ WSI を一度に学習するのはデータ数、パラメータサイズどちらの面においても現実的ではない。また、癌などの異常部位というのは WSI においてわずか領域にのみ現れることも多いため、基本的には WSI を大量の画像パッチに分割し、それぞれを識別した結果を統合することで癌の有無を判定することが多い。個々の画像パッチを識別するのは、SVM, Convolutional Neural Network などが用いられる。

2.1.3 特徴抽出

機械学習において特徴抽出は重要な役割を持つ。従来は hand-crafted による特徴量 (SIFT, HLAC, ToDo: ちょっと調べる) が使われていた。また、テクスチャとしての性質を持つことからテクスチャ解析における手法をしばしば利用することがある。しかし、近年では一般画像認識の成果を受けて CNN を利用して同時に学習することが多い。また、pretrained model による特量を利用することもある。

2.1.4 能動学習の適用

(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

人手による前処理 + uncertainty sampling のみ [?] class balancing + query-by-committee[?], logistic regression + variance reduction [?], 仮説空間の縮小を目的関数にすることで劣モジュラ最適化の枠組みを利用、似たデータが多いことを利用して k-means でクラスタリング [?]、

2.2 深層能動学習

この節では、深層能動学習の先行研究について述べる。

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[?] 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.

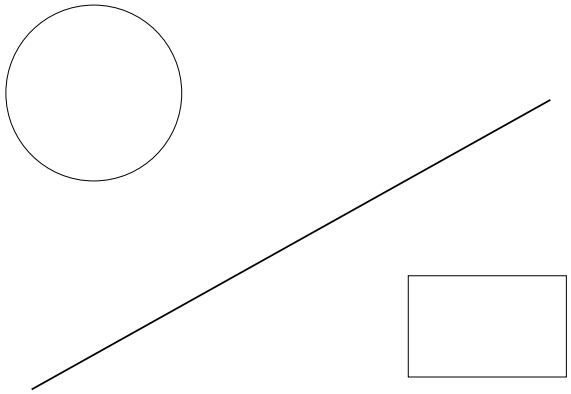

図 4.1. 図のタイトル

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．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．

謝辞

TODO

付録 A

ソースコード

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