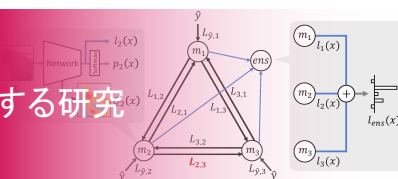


2021年度 藤吉研究室 修士論文発表 アブストラクト

Deep Learning Ensemble Learning Knowledge Distillation

多様な知識蒸留による深層学習モデルのアンサンブル学習に関する研究

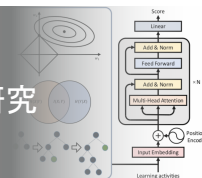
岡本 直樹



Deep Learning Learning Analytics

Learning Analyticsにおける学習活動を用いたTransformerによる成績予測に関する研究

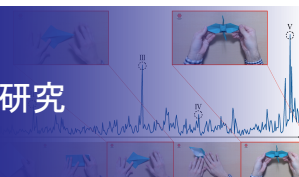
佐竹 央義



Deep Learning Skill Assessment Explainable AI

複数の深層学習モデルによるスキル優劣判定とその判断根拠に関する研究

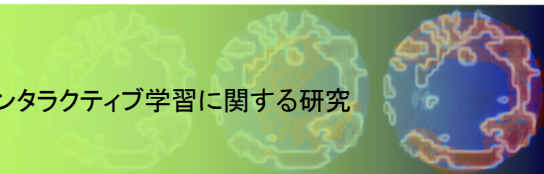
高田 雅之



Deep Learning Fine Tuning Explainable AI

AIから学ぶ : エキスパートの知見を導入したDNNモデルを教師とするインタラクティブ学習に関する研究

服部 幸平



1. はじめに

複数の深層学習モデルによるアンサンブル学習では、各ネットワークの出力分布の平均から推論を行うことで、1モデルの推論と比べて認識性能が向上する。このとき、各ネットワークの出力分布間に多様性が生まれると、アンサンブルの効果が高くなると考えられる。しかし、ネットワーク間で出力分布を離すように学習すると、出力分布の多様化が期待できるが、性能低下が生じる。

そこで本研究では、性能低下を引き起こすことなく多様性を得るために Attention map に着目し、出力分布と Attention map を知識とした知識蒸留を複数のネットワーク間で行う。そのために、従来の知識蒸留にネットワーク間の知識を離す設計を導入する。これにより、アンサンブル学習に適した多様な知識蒸留を実現できる。また、知識蒸留を用いたアンサンブル学習にグラフ表現 [2] を導入することで、多様なアンサンブル学習方法を獲得可能とする。

2. 知識蒸留

知識蒸留 [1] は、ネットワークが獲得した知識を他のネットワークが模倣することで、知識を転移する学習手法である。知識の転移方向として一方向や双方向、知識として確率分布や特徴マップ、Attention map など様々な方法が提案されている。南ら [2] は、知識蒸留をグラフで表現し、知識の転移方法をハイパーパラメータとして最適化することで、従来の様々な知識蒸留方法を内包し、その中から効果的な知識蒸留方法を自動設計可能な手法を提案した。従来の知識の転移方法は、ネットワークの認識性能が向上する一方、ネットワーク間で多様性が生まれないため、知識蒸留を行わないネットワークと比べて同程度のアンサンブル性能となる。

3. 提案手法

本研究では、ネットワークの多様性を獲得するために、確率分布と Attention map の観点から各ネットワークが異なる知識を獲得できるような知識蒸留を行う。また、グラフ表現 [2] をアンサンブル学習に拡張し、グラフ構造をハイパーパラメータ最適化することでモデル数に応じた効果的なアンサンブル学習方法を自動設計する。

3.1 多様な知識蒸留を用いたアンサンブル学習

アンサンブル学習のための知識蒸留として、知識を近づける損失と離す損失を設計する。損失設計は、最小化問題として学習を行うために近づける・離す場合で異なる設計を用いる。この時、知識の転移先をターゲットネットワーク、知識の元をソースネットワークとする。

確率分布を近づける場合は KL-divergence、離す場合はコサイン類似度を利用する。KL-divergence を用いた損失関数は次のように定義する。

$$L_p(x) = \sum_{c=1}^C p_s^c(x) \log \frac{p_s^c(x)}{p_t^c(x)} \quad (1)$$

ここで、 C はクラス数、 x は入力サンプル、 $p_s^c(\cdot)$ はソースネットワークのクラス c の確率値、 $p_t^c(\cdot)$ はターゲットネットワークのクラス c の確率値である。コサイン類似度を用いた損失関数は次のように定義する。

$$L_p(x) = \frac{p_s(x)}{\|p_s(x)\|_2} \cdot \frac{p_t(x)}{\|p_t(x)\|_2} \quad (2)$$

ここで、 $p_s(\cdot)$ はソースネットワークの確率分布、 $p_t(\cdot)$ はターゲットネットワークの確率分布である。

Attention map は、入力サンプルの認識に有効な領域に強く反応する。対象物体の大きさはサンプルごとに異なるため、対象物体の異なる箇所に強く反応しているのに関わらず、類似度が高くなる場合がある。そこで、Attention map のクロップを行う。ソースネットワークの Attention

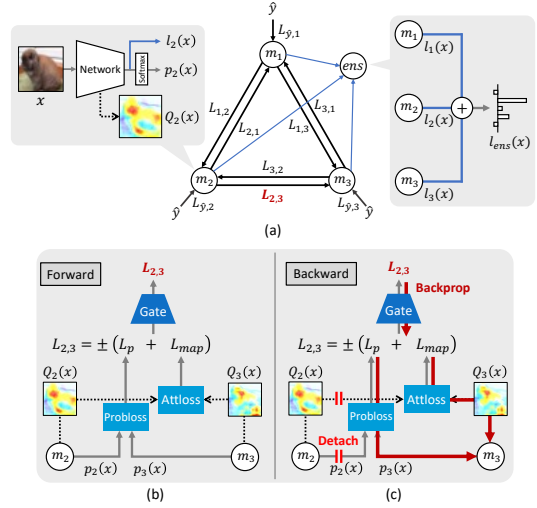


図 1: グラフ表現を導入した多様な知識蒸留を用いたアンサンブル学習

map において最も値の高い位置を中心に正方形のクロップを行い、ターゲットネットワークはソースネットワークと同じ位置をクロップする。複数のサイズでクロップを行い、それぞれのサイズにおける損失値の平均値を最終的な損失値とする。Attention map を近づける場合は平均二乗誤差、離す場合はコサイン類似度を利用する。平均二乗誤差を用いた損失関数は次のように定義する。

$$L_{map}(x) = \frac{1}{K} \sum_{k=1}^K \left(\frac{Q_s^k(x)}{\|Q_s^k(x)\|_2} - \frac{Q_t^k(x)}{\|Q_t^k(x)\|_2} \right)^2 \quad (3)$$

ここで、 K はクロップ数、 $Q_s^k(\cdot)$ はクロップしたソースネットワークの Attention map、 $Q_t^k(\cdot)$ はクロップしたターゲットネットワークの Attention map である。コサイン類似度を用いた損失関数は次のように定義する。

$$L_{map}(x) = \frac{1}{K} \sum_{k=1}^K \frac{Q_s^k(x)}{\|Q_s^k(x)\|_2} \cdot \frac{Q_t^k(x)}{\|Q_t^k(x)\|_2} \quad (4)$$

3.2 アンサンブル学習のためのグラフ表現

アンサンブル学習にとって効果的な知識蒸留方法を自動設計するために、グラフ表現 [2] を利用する。グラフ表現を導入した知識蒸留を用いたアンサンブル学習を図 1(a) に示す。グラフはノードとエッジで構成される。ノードは、ネットワークを表すネットワークノードとアンサンブルを行うアンサンブルノードを定義する。エッジは、損失計算を表し、ネットワークノード間のエッジは知識蒸留、ネットワークノードとラベル $\hat{y}$ 間のエッジは教師ラベルを用いた交差エントロピー損失を表す。

アンサンブルノードは、全てのネットワークノードの出力を用いたアンサンブルを行う。アンサンブルノードにおける処理は次のように定義する。

$$l_{ens} = \frac{1}{M} \sum_{m=1}^M l_m(x) \quad (5)$$

ここで M はネットワークノード数、 $l_m(\cdot)$ はネットワークノードの logit 関数、 x は入力サンプルである。

ネットワークノード間のエッジは、ノード間の知識蒸留を行う。エッジでは、図 1(b) のように知識蒸留の損失値に対してゲート [2] を用いて重み付けを行うことで、知識蒸留を制御する。知識蒸留の損失計算は次のように定義する。

$$L_{s,t}(x) = G_{s,t}(L_p(x) + L_{map}(x)) \quad (6)$$

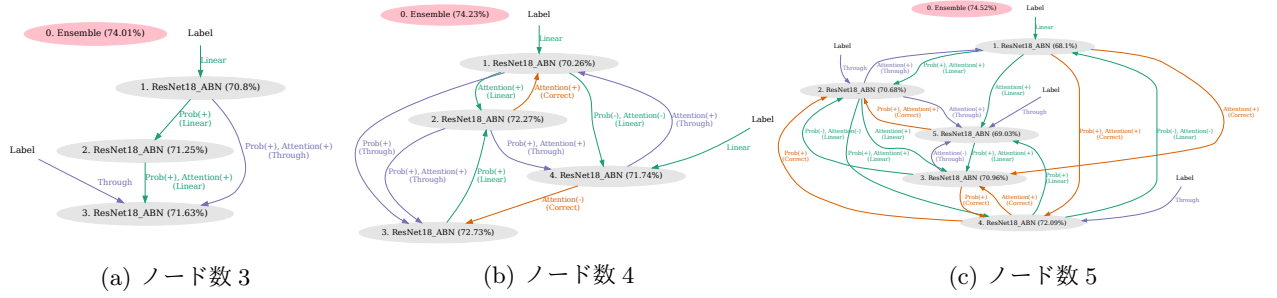


図 2: 最適化後のグラフ構造: 赤色のノードはアンサンブルノード, 灰色のノードはネットワークノード, “Label” は教師ラベル, “Prob” と “Attention” は損失設計の種類を表す. エッジの色はゲートの種類を表す.

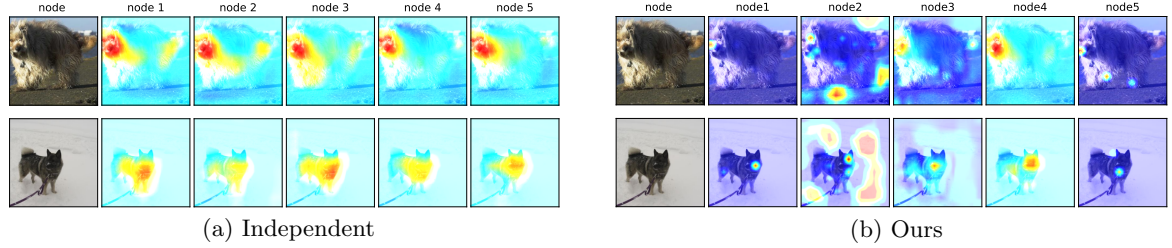


図 3: Attention map の可視化

表 1: 従来手法との精度比較

手法	ネットワーク	精度 [%]	
		ノード平均	アンサンブル
Independent	ABN × 3	68.04 ± 0.28	71.41 ± 0.34
DML	ABN × 3	70.50 ± 0.26	72.08 ± 0.42
Ours	ABN × 3	70.95 ± 0.16	73.41 ± 0.30
Independent	ABN × 4	68.30 ± 0.27	72.06 ± 0.53
DML	ABN × 4	71.50 ± 0.31	72.87 ± 0.29
Ours	ABN × 4	71.46 ± 0.22	74.16 ± 0.22
Independent	ABN × 5	68.24 ± 0.26	72.32 ± 0.18
DML	ABN × 5	71.15 ± 0.28	72.50 ± 0.16
Ours	ABN × 5	70.23 ± 0.33	74.14 ± 0.50

表 2: 異なるデータセットにおけるアンサンブル精度

学習・評価用 データセット	最適化用 データセット	ノード数		
		3	4	5
CUB-200-2011	Stanford Dogs	71.82	73.03	72.13
CUB-200-2011	CUB-200-2011	74.17	72.22	74.05
Stanford Cars	Stanford Dogs	89.94	89.98	90.41
Stanford Cars	Stanford Cars	90.04	89.57	90.73

ここで $G_{s,t}(\cdot)$ はゲートを表す. 知識蒸留の損失から計算された勾配は, エッジの向きによって勾配を伝えるノードが変化する. $L_{2,3}$ の場合は図 1(c) のように, ノード m_3 にのみ勾配を伝えるためにノード m_2 側の計算グラフをカットすることで, ノード m_2 からノード m_3 への知識蒸留が行われる.

グラフのハイパーパラメータは, ネットワークノード間のエッジにおける損失設計, 各エッジの損失値に適用されるゲートである. 損失設計は, 確率分布を近づける・離す, Attention map を近づける・離す, 確率分布と Attention map を同時に近づける・同時に離すの計 6 種類とする. ゲートは, 従来法 [2] で提案された 4 種類とする. ハイパーパラメータ最適化は, ランダムサーチと枝刈りを用いて 6,000 組のグラフ構造を評価する.

4. 評価実験

ネットワークノード数を 3 から 5 としてハイパーパラメータ最適化をしたグラフの評価を行う. ネットワークは, Attention Branch Network(ABN) [3] を用いる. データセットは, 詳細画像識別問題である Stanford Dogs, Stanford Cars, Caltech-USCD Birds-200-2011(CUB-200-2011) を用いる. 損失計算時に Attention map は, 3×3 , 7×7 , 11×11 のサイズにクロップする.

4.1 最適化したグラフの評価

Stanford Dogs を用いて最適化を行ったグラフ構造を図 2, Stanford Dogs における従来手法との比較を表 1 に示す. Independent は個別に学習を行ったネットワーク, DML は従来の知識蒸留を行ったネットワーク, Ours は最適化したグラフの結果である. Ours は, Independent や DML を

上回るアンサンブル精度を達成した.

Independent の Attention map を図 3(a), 図 2(c) のグラフにおける Ours の Attention map を図 3(b) に示す. Ours は, 多様な知識蒸留によってネットワークごとに異なる注視領域を獲得したと言える.

4.2 グラフ構造の汎化性

Stanford Dogs を用いて最適化したグラフを最適化時と異なるデータセットを用いて評価する. CUB-200-2011, Stanford Cars における評価結果を表 2 に示す. 最適化に用いたデータセットによらず同程度のアンサンブル精度となっていることから, 最適化したグラフ構造に汎化性があると言える.

5. おわりに

本研究では, アンサンブル学習のための多様な知識蒸留とグラフ表現を提案した. 評価実験により, 従来の知識蒸留やアンサンブル学習を超える性能を発揮するアンサンブル学習方法を獲得したことを確認した. 今後は, 獲得したグラフ構造の分析やグラフの最適化方法について検討する.

参考文献

- [1] G. Hinton, *et al.*, “Distilling the knowledge in a neural network”, NeurIPS workshop, 2015.
- [2] S. Minami, *et al.*, “Knowledge Transfer Graph For Deep Collaborative Learning”, ACCV, 2020.
- [3] H. Fukui, *et al.*, “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”, CVPR, 2019.

研究業績

- [1] 岡本直樹 等, “知識蒸留グラフによるアンサンブル学習”, 画像の認識・理解シンポジウム, 2020.
- [2] 岡本直樹 等, “詳細画像識別における知識蒸留グラフによるアンサンブル学習”, 画像の認識・理解シンポジウム, 2021.

1. はじめに

Learning Analytics (LA) とは、教育分野において学習者や指導者の活動などのデータを分析し、有益な情報の取得及び利用を目的としたデータサイエンスである。教育現場でのデジタル学習環境やオンライン教育システムの普及により、膨大なユーザの学習活動データの収集が可能になっている。そのデータを分析することで、ユーザの学習パターンから成績を予測し、ユーザへのサポートに活かすことができる。

デジタル学習プラットフォームで収集されたデータから学生の成績を予測した研究例として、Okubo らの研究がある [1]。この研究では大学の講義時間内における学習活動から特徴量を作成し、時系列情報を考慮できる深層学習モデルである Recurrent Neural Network (RNN) によって成績の評価レベルを予測している。しかし、この手法では自宅学習など講義時間外の学習活動の影響を考慮できず、また RNN ではどのような活動が予測に影響を与えているかといった分析は困難である。そこで本研究では、多数の特徴量からの特徴量選択とアテンションを用いた深層学習手法である Transformer [2] を用いたモデルによって、学生の学習活動データからの成績予測を行う。

2. 関連研究

2.1 特徴量選択

機械学習における特徴量選択手法には大別して、統計的手法によって特徴量の評価を行って選択する Filter Method、特徴量を入れ替えながらモデルの学習を行って最適な特徴量を選択する Wrapper Method、機械学習における重要度や重みなどの値で選択を行う Embedded Method がある。Filter Method には特徴量の分散のサイズや目的変数との相関係数を利用した選択、Embedded Method は決定木モデルの feature importance を利用した選択などがある。

2.2 Transformer

Transformer は、自然言語処理分野で発表された、再帰や畳み込みを用いずに時系列情報を考慮できるようにしたエンコーダ-デコーダモデルである。Transformer のエンコーダ構造として、まず入力シーケンスの各要素に位置情報を付加する Positional Encoding がある。ブロックは、複数のヘッドそれぞれで Self-Attention を計算する Multi-Head Attention、シーケンスごとに位置情報を整える Feed-Forward Network、残差結合と正規化を行う Add & Norm 層の 3 種類のサブレイヤーで構成される。Transformer では Self-Attention によって時系列情報などの入力シーケンスの関係性を考慮できる。Self-Attention は入力シーケンスごとに作成した 3 つのベクトル Query, Key, Value の内積によって求められる。

3. 提案手法

本研究では、講義時間外の学習活動を含めた多数の特徴量から、特徴量選択によって予測モデルの学習に有効な特徴量を選択し、モデルの学習に用いることを提案する。学習活動を表現する多数の特徴量を作成し使用することで、モデルはより詳細な学習活動のパターンを把握できると考えられる。しかし、特徴量全てを使用すると計算コストが膨大になるため、予測に有効な特徴量を選択することで効率的なモデルの学習を行う。提案手法による成績予測の処理を図 1 に示す。

3.1 特徴量選択

特徴量選択手法には、Filter Method として特徴量の分散の大きさ、成績との Mutual information, Embedded Method として Lasso 回帰の重み, LightGBM の feature importance の計 4 種類の手法を用いる。なお Wrapper Method については、選択するための学習自体に膨大な計算コストがかかるため今回は使用しない。

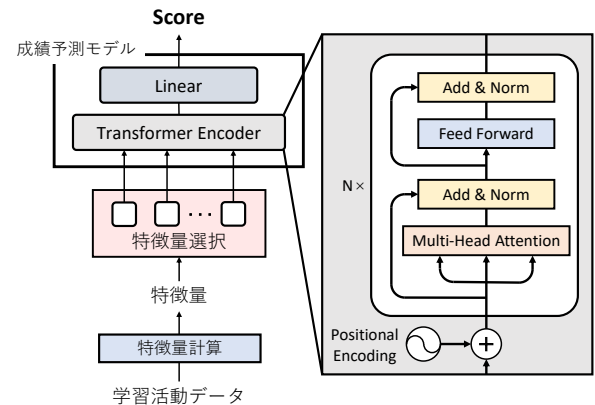


図 1：提案手法による予測までの処理

3.2 Transformer を用いた予測モデル

予測モデルには、Transformer のエンコーダ構造を利用したモデルを用いる。これにより、入力同士の関係性を考慮でき、また入力について Self-Attention による分析が可能となる。特徴量選択により選択された特徴ベクトルを入力として Transformer のエンコーダによって潜在的な特徴を学習し、全結合層によって成績を出力する。

4. 提案手法による予測実験

提案手法を用いて、4 つの成績予測実験を行う。1 つ目は、特徴量選択手法の有効性評価である。4 種類の特徴量選択手法を用いた際の予測モデルでの予測精度を比較する。2 つ目は、講義時間外の学習活動データの有効性評価である。講義時間内の学習活動データのみから作成した特徴量と、講義時間外を含めた全時間帯の学習活動データから作成した特徴量とで成績予測を行う。3 つ目は Transformer を利用したモデルと、LA 分野の成績予測タスクにおいて用いられている従来手法のモデルとの比較である。従来手法として、LSTM, MLP, 重回帰分析を用いる。4 つ目は成績の早期予測である。一連の講義の進行途中を想定し、使用する特徴量を制限した成績予測を行う。

4.1 データセット

実験には、九州大学で行われた情報系科目の一連の講義における学習活動データから作成した特徴量を使用する。講義は 8 回行われ、各回で異なる講義資料が電子教材として使用された。また、全講義終了後には 100 点満点のテストが実施された。本実験では、テストの点数を講義における成績として扱う。受講した学生 1300 人のうち、学習データとして 1,200 人分、評価データとして 100 人分のデータを使用する。

学習活動データは、M2B 学習支援システム [3] によって収集された。M2B 学習支援システムは九州大学で利用されているデジタル学習プラットフォームで、学生の電子教材の操作ログなどの学習活動データを記録している。操作ログには、学生が取ったアクションの種類、アクション対象のページ番号、タイムスタンプなどが記録されている。アクションの種類には、ページ送りの「NEXT」や戻りの「PREV」、文字にハイライトを追加する「ADD MARKER」など計 16 種類がある。また、電子教材の使用は講義時間外も可能なため、自宅学習など講義時間外の操作ログも収集される。

特徴量として、講義時間外を含めた全時間帯の学習活動データから各ページにおける各アクション回数と各ページの閲覧時間の特徴量を作成し、また講義時間内の学習活動データのみから 5 分ごとの各アクション回数の特徴量を作成した。これらの特徴量は合計で 9,555 次元となっており、各次元で正規化を行った。

4.2 特徴量選択手法の有効性評価

特徴量選択手法として分散の大きさ、Mutual information, Lasso 回帰, feature importance を用いて各手法の比較を行う。9,555 次元のうち 128 次元を選択し、選択した特徴ベクトルを用いて予測モデルの学習と評価を行う。

200 epoch の学習を 5 回ずつ行ったときの、評価データに対するテスト点数予測値の平均平方二乗誤差 (RMSE) の平均を表 1 に示す。4 つの手法の中で、Lasso 回帰によって選択された特徴量を用いた場合が最も RMSE が小さかった。このことから、Lasso 回帰が最も予測に有効な特徴量を選択できているといえる。

表 1：各特徴量選択手法による予測値の RMSE

特徴量選択手法	RMSE
分散の大きさ	9.20±0.43
Mutual information	9.40±0.67
Lasso 回帰	8.90±0.45
feature importance	9.10±0.38

また、各特徴量選択手法でどのような特徴量が選択されたか調査した。図 2 に選択された各アクションに関する特徴量数のグラフを示す。最も予測精度が低かった Mutual information では、ページ送りである「NEXT」のアクションに関する特徴量が集中的に選択されており、16 種類あるうちの 6 種類のアクションからしか特徴量が選択されていなかった。それに対し最も精度の高かった Lasso 回帰は、他手法では選択されなかった 6 種類を含めた 14 種類のアクションから特徴量が選択されていた。予測の際に様々なアクションの特徴量を用いたことで学習活動パターンをより細かく把握でき、精度が高くなったのではないかと考えられる。

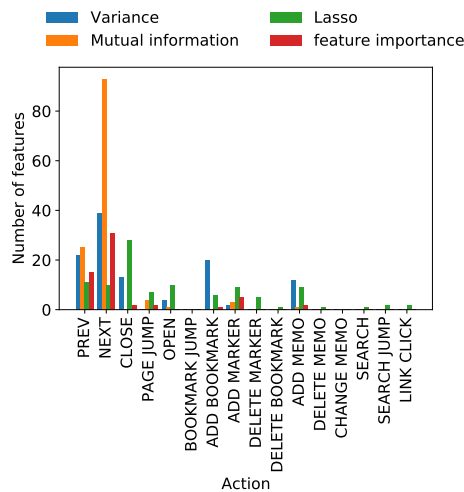


図 2：各手法で選択されたアクションに関する特徴量

4.3 講義時間外の学習活動データの有効性評価

講義時間外を含めた特徴量と、講義時間内のみの特徴量による予測との比較を行う。各講義内で行った 16 種類のアクションの回数を講義時間内のみの特徴量とし、計 128 次元を Transformer モデルに入力する。講義時間外を含めた特徴量としては、4.2 節と同じく 9555 次元の特徴量から特徴量選択によって 128 次元の特徴ベクトルを選択する。選択手法には Lasso 回帰を用いる。

5 回ずつ学習を行ったときの、予測値の RMSE の平均を表 2 に示す。講義時間内のみの場合と比較して、講義時間外を含めた特徴量を用いた場合に RMSE が低くなった。結果から、講義時間外の学習活動データが予測精度向上に有効であると考えられる。

表 2：各特徴量選択手法による予測値の RMSE

特徴量	RMSE
講義時間内のみ	9.10±0.69
講義時間内+講義時間外	8.90±0.45

4.4 従来手法との比較

成績予測タスクにおいて主に用いられる LSTM, MLP, 重回帰分析を用いた予測との比較を行う。これらのモデルの特徴量には、16 種類のアクションを各講義内で行った回数の、128 次元の特徴ベクトルを使用する。

5 回ずつ学習を行ったときの、予測値の RMSE の平均を表 3 に示す。従来手法と比較して Transformer を利用したモデルの RMSE が最も低く、高精度な予測ができることが分かった。

表 3：各モデルによる予測値の RMSE

予測モデル	RMSE
LSTM	9.34±0.52
MLP	9.37±0.53
重回帰分析	10.02±0.52
Transformer	8.90±0.45

4.5 成績の早期予測

予測結果を教育現場にフィードバックすることを考えたとき、より早期の段階で成績を予測したり、その時々での予測結果を示すことで、適切な支援に活かせることが望ましい。そこで、一連の講義の進行途中を想定した、成績の早期予測実験を行う。作成した特徴量 9,555 次元のうち、特定の講義回までに使用された講義資料に関する特徴量から特徴量選択を行い、Transformer モデルの学習および評価に使用する。例えば、4 回目の講義時点での予測では 1～4 回目で使用された講義資料に関する特徴量から選択を行う。

各特徴量選択手法を用いて学習を行ったときの 1～8 回目までの RMSE の推移を図 3 に示す。いずれの手法を用いた場合でも概ね講義が進むにつれて RMSE が小さくなった。また、Lasso 回帰は全体を通じて RMSE が最も小さいか僅差で 2 番目に小さく、早期予測においても高い精度であることが分かった。

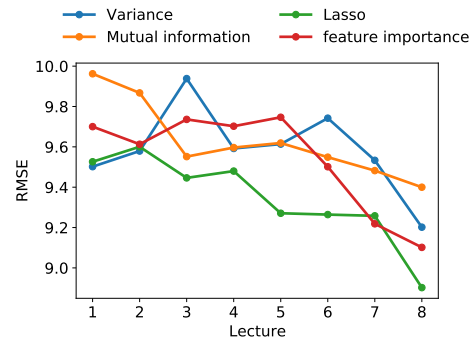


図 3：各特徴量選択手法での RMSE の推移

5. おわりに

本研究では、特徴量選択と Transformer を利用したモデルによるテスト点数の予測を行った。特徴量選択手法としては Lasso 回帰が最も精度が高くなった。また、成績予測タスクにおける従来手法と比較して、Transformer を利用したモデルは高精度の予測結果が得られた。今後は、作成する特徴量の改善とより効果的な特徴量選択を検討する。

参考文献

- [1] F.Okubo, *et al.*, “A Neural Network Approach for Students’Performance Prediction”, LAK17, 2017
- [2] A.Vaswani, *et al.*, “Attention is all you need”, NIPS, 2017
- [3] H.Ogata, *et al.*, “M2B System: A Digital Learning Platform for Traditional Classrooms in University”, LAK17, 2017

研究業績

- [1] A.satake, *et al.*, “Performance prediction and importance analysis using Transformer”, ICCE, 2021 (他 1 件)

1. はじめに

スキル優劣判定は、動画内の動作の優劣判定を行うタスクであり、これまでに Rank-aware Attention Networks (RAAN) [1] が提案されている。RAAN の学習には、優劣の2値ラベルが付与された2つの動画を入力して、Ranking Loss により重みの更新を行う。このとき、動画を K 個のセグメントに分割し、全てのセグメントが正解になるよう学習する。優劣の2値ラベルは動画全体を見てラベルが付与されており、動画内に含まれる全てのセグメントが該当するわけではない。しかし RAAN は、優劣の差が無いセグメントに対して全体に付与されたラベルを使用するため、誤った優劣判定を誘発するという問題がある。本研究では、この問題を解決するアプローチとして、Multiple Instance Learning (MIL) を用いて、優劣の差があるセグメントに着目したスキル優劣判定法を提案する。提案手法は、動画をバッグ、動画から切り出したセグメントをインスタンスとし、バッグのみにラベルを付与して優劣判定を行う。これにより、優劣の差があるインスタンスのみに着目した学習が可能となる。また、優れた動作と劣った動作を個別に判定することで、それぞれの時間情報に関する判断根拠の獲得を目指す。

2. 関連研究

スキル優劣判定は、スキルを評価する手法であるため、医療やスポーツ等を対象とした研究が行われている。動画共有サービスによって様々なスキルが含まれた動画が溢れる現在では、様々な動作に使用できるスキル優劣判定の研究が必要となる。様々な動作に使用できるスキル優劣判定手法の中で Rank-aware Attention Networks (RAAN) [1] は、比較的長い動画に対してスキル優劣判定を行う手法である。この手法では、Temporal Attention を導入し、動画内における優れたセグメントと劣ったセグメントを識別することができる。

3. 提案手法

本研究では、優劣の差があるセグメントに着目するためのスキル優劣判定手法である Skill MIL を提案する。

3.1 Skill MIL

提案手法の構造を図1に示す。まず、動画を K 個のセグメントに分割する。それぞれのセグメントから連続する16フレームをランダムに選択する。提案手法では MIL を導入するために、1つのセグメントをインスタンス k とし、 K 個のインスタンスの集合体をバッグ $\beta = \{k_x | x = 1, \dots, K\}$ と定義する。インスタンスにはラベルを付与せず、バッグのみにラベルを付与して学習を行う。入力、優れたスキルを表す動画を p_i と、劣ったスキルを表す動画を p_j のペアとする。動画は $p_i > p_j$ の関係であり、 p_i から構成されるバッグをポジティブバッグ β^p 、 p_j から構成されるバッグをネガティブバッグ β^n とする。 β^p 、 β^n に含まれるインスタンス k_x^p 、 k_x^n を Feature extractor に入力する。Feature extractor からの出力を、それぞれ Superior ranking module (SRM) と Inferior ranking module (IRM) に入力する。その後、SRM と IRM の出力を演算することで最終的なスコアを獲得し、優劣判定を行う。

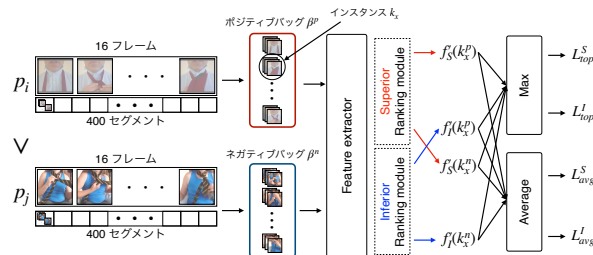


図1：提案手法の構造

3.2 Ranking module

SRM と IRM の内部構造は同一であり、その構造を図2に示す。各モジュールは、3層の全結合層と Attention sub module

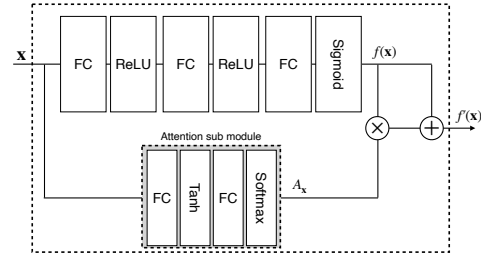


図2：Ranking module の構造

module で構成する。Feature extractor の出力 $\mathbf{x}$ を入力としたとき、Attention sub module はインスタンスの重要度 $A_{\mathbf{x}} = \{a_k | k = 1, \dots, K\}$ を出力する。各インスタンスの重要度 a_k の算出方法を式 (1) に示す。

$$a_k = \frac{\exp\{\mathbf{w}^T \tanh(\mathbf{V}\mathbf{x}_k^T)\}}{\sum_{j=1}^K \exp\{\mathbf{w}^T \tanh(\mathbf{V}\mathbf{x}_j^T)\}} \quad (1)$$

ここで、 $\mathbf{w}$, $\mathbf{V}$ は学習するパラメータを表す。獲得したインスタンスの重要度 $A_{\mathbf{x}}$ は、式 (2) に示すように3層の全結合層からの出力である特徴マップ $f(\mathbf{x})$ に重み付けとして乗算し、重み付け前の特徴を加算する。これにより、特徴マップの消失を抑制し、インスタンスの重要度を優劣判定に反映させることができる。

$$f'(\mathbf{x}) = f(\mathbf{x}) + (A_{\mathbf{x}} \cdot f(\mathbf{x})) \quad (2)$$

3.3 損失関数と学習

Skill MIL で使用する損失関数について説明する。SRM の損失関数 L_{top}^S と L_{avg}^S は、 $\beta^p > \beta^n$ の関係になるよう学習することで、動画内の優れた動作に焦点を当てた学習ができる。 L_{top} は、バッグ内の最大値を比較する損失関数であり、 L_{avg} は、バッグ内の平均値を比較する損失関数である。 L_{top}^S と L_{avg}^S を、式 (3), (4) に示す。

$$L_{top}^S = \max(0, 1 - \max_{k_x^p \in \beta^p} f'_S(k_x^p) + \max_{k_x^n \in \beta^n} f'_S(k_x^n)) \quad (3)$$

$$L_{avg}^S = \max(0, 1 - \frac{1}{\beta} ((\sum_{\beta} f'_S(k_x^p)) + (\sum_{\beta} f'_S(k_x^n)))) \quad (4)$$

ここで、 f'_S は SRM の出力を表す。 L_{top}^S は、 β^n より β^p の方が最大値が高くなるよう学習し、 L_{avg}^S は、 β^n より β^p の方が平均値が高くなるよう学習する。

一方で、IRM の損失関数 L_{top}^I と L_{avg}^I は、 $\beta^n > \beta^p$ の関係になるよう学習することで、動画内の劣った動作に焦点を当てた学習が可能となる。 L_{top}^I と L_{avg}^I を、式 (5), (6) に示す。

$$L_{top}^I = \max(0, 1 - \max_{k_x^n \in \beta^n} f'_I(k_x^n) + \max_{k_x^p \in \beta^p} f'_I(k_x^p)) \quad (5)$$

$$L_{avg}^I = \max(0, 1 - \frac{1}{\beta} ((\sum_{\beta} f'_I(k_x^n)) + (\sum_{\beta} f'_I(k_x^p)))) \quad (6)$$

ここで、 f'_I は IRM の出力を表す。 L_{top}^I は、 β^p より β^n の方が最大値が高くなるよう学習し、 L_{avg}^I は、 β^p より β^n の方が平均値が高くなるよう学習する。

Skill MIL 全体の損失関数を、式 (7) に示す。

$$L = L_{top}^S + L_{top}^I + \lambda L_{avg}^S + \lambda L_{avg}^I \quad (7)$$

ここで、 λ は寄与率を調整するための係数であり、 $\lambda = 0.5$ を用いる。

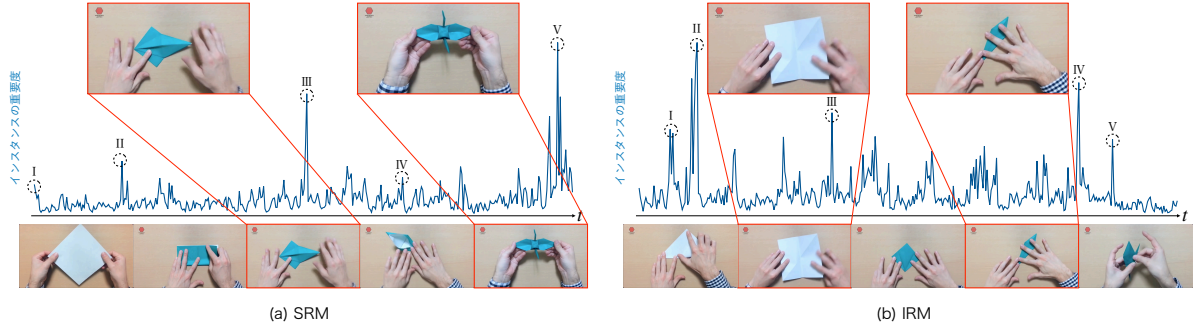


図 3：高いスキルレベルの動画に対するインスタンスの重要度 (Origami タスク)

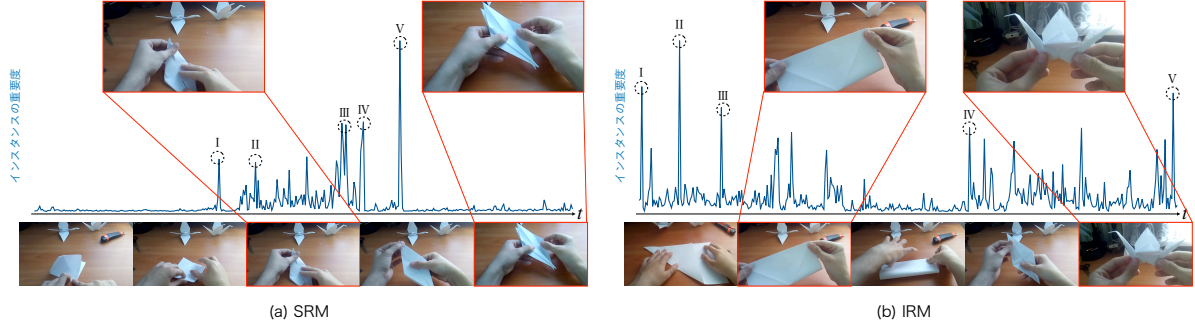


図 4：低いスキルレベルの動画に対するインスタンスの重要度 (Origami タスク)

3.4 優劣判定

本研究では、対応するインスタンスの優秀かどうかを表すスコア $score_{high}$ と、劣悪かどうかを表すスコア $score_{low}$ を比較することでスキル優劣判定を行う。それぞれ式 (8), (9) により求める。

$$score_{high} = \max_{k_x^p \in \beta^p} f'_S(k_x^p) + \max_{k_x^n \in \beta^n} f'_I(k_x^n) \quad (8)$$

$$score_{low} = \max_{k_x^n \in \beta^n} f'_S(k_x^n) + \max_{k_x^p \in \beta^p} f'_I(k_x^p) \quad (9)$$

$score_{high}$ と $score_{low}$ を比較し、スコアの高い方を判定結果とする。

4. 評価実験

提案手法の有効性調査を行う。比較する手法には RAAN を用いる。データセットには BEST Dataset を使用する。Best Dataset は、YouTube から取得した動画で構成され、Apply Eyeliner, Braid Hair, Origami, Scramble Eggs, Tie Tie の 5 つのサブセットで構成されるデータセットである。RAAN と実験条件を揃えるため、公開されている Best Dataset の I3D 特徴を使用する。本 I3D 特徴は、Kinetics 400 データセットで事前学習したモデルを使用して抽出されている。また、使用する際は過学習を抑制するためにノイズ $\{z|0 < z < 1\}$ を I3D 特徴に加算してから実験を行う。最適化手法には Adam を使い、学習率は 0.0001、バッチサイズを 128、 $K = 400$ とし、試行回数を 2000 エポックとした。

4.1 実験結果

表 1：従来手法との比較 [%]

Method	Apply Eyeliner	Braid Hair	Origami	Scramble Eggs	Tie Tie	Average
RAAN	84.3	74.4	78.8	84.4	87.1	81.8
Ours	83.4	75.2	80.7	90.3	87.1	83.3

従来手法である RAAN と比較した結果を表 1 に示す。表 1 より、提案手法は平均して 1.5pt 従来手法より精度が向上した。特に Scramble Eggs タスクでは、5.9pt 向上した。次に、インスタンスの重要度の可視化例を図 3, 4 に示す。図 3 (a) と図 4 (a) に示すように、SRM では、動画の中盤部分を強く重要視していることが確認できる。これは中盤部分の作業が折り鶴の完成度に大きく影響を与えている為だと考えられる。また、図 3 (a) では、中盤部分の

みではなく動画の終盤に対しても強く重要視している。これはこの動画の終盤で、綺麗に折られた折り鶴をカメラに写している (図中 V) ためだと考えられる。一方で、IRM では、図 3 (b) と図 4 (b) に示すように動画の序盤や終盤を強く重要視していることが確認できる。スキルレベルの低い動画は、折り直す動作を複数回行うことが多い。序盤を重要視する原因は、複数回行う折り目を付ける動作を折り直す動作と判定しているからだと考えられる。また終盤を重要視する原因は、完成した折り鶴を見せている動画が多いからだと考えられる。

4.2 インスタンスの重要度の有効性

表 2：インスタンスの重要度 反転時の精度変化 [%]

Inversion	Apply Eyeliner	Braid Hair	Origami	Scramble Eggs	Tie Tie	Average
before	83.4	75.2	80.7	90.3	87.1	83.3
after	53.6	63.9	52.8	51.1	78.1	59.9

提案手法で獲得したインスタンスの重要度 A_x の有効性を評価する。評価方法として、インスタンスの重要度を反転させて推論を行う。反転する前と反転した後での精度を比較することで精度がどれだけ低下するかを調査する。表 2 にインスタンスの重要度の反転時の精度変化を示す。表 2 より、インスタンスの重要度の反転により大幅に精度が低下することが確認できた。特に、Scramble Eggs タスクでは 39.2pt の低下が確認できた。よって、スキル優劣判定に有効なインスタンスの重要度をえられる事が確認できる。

5. おわりに

本研究では Skill MIL を提案し、優劣の差があるセグメントに着目する学習を行い従来手法より高精度を獲得した。また、インスタンスの重要度を獲得し、ネットワークに適用することで、精度の向上と時間情報の判断根拠が獲得可能なこと示した。今後は、空間情報に対する重要度を獲得し、更なる精度向上を検討する。

参考文献

- [1] H. Doughty. *et al.*: “The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos”, CVPR. 2019.

研究業績

- [1] 高田雅之, 等, “Attention Pairwise Ranking によるスキル優劣判定における視覚的説明と高精度化”, 第 23 回 画像の認識・理解シンポジウム, 2020.

AI から学ぶ：
エキスパートの知見を導入した
DNN モデルを教師とするインタラクティブ学習に関する研究
TR20011 服部 幸平

指導教授：藤吉 弘亘

1. はじめに

Attention Branch Network (ABN) [1] は、深層学習の判断根拠を人が解釈できるよう、推論時における注視領域をアテンションマップとして可視化する視覚的説明手法である。さらに ABN は、アテンションマップを特徴マップに重みづけをすることで認識性能の向上を実現した。三津原ら [2] は、アテンションマップをエキスパートの知見に基づいて修正し、モデルをファインチューニングすることで視覚的説明性と認識性能をさらに向上させた。

本研究では、視覚的説明を人が解釈するにとどまらず、AI から人が学ぶ Human-in-the-loop を実現することを目的とする。そこで、エキスパートの知見を組み込んだ深層学習モデルを教師役として、学習者がその判断根拠をインタラクティブに学習する方法を提案する。提案手法は、学習者が編集したアテンションマップを利用して、エキスパートの知見を組み込んだ ABN が推論する。この時の推論したスコアが高くなるように、学習者によるアテンションマップの修正と ABN による推論を繰り返す。これにより、ABN に組み込んだエキスパートの判断根拠である注視領域をインタラクティブに学習することが可能となり、高い教育効果を実現する教育アプリとなる。教育アプリを用いて、眼底画像の疾患判定を対象とした被験者による評価実験を行い、提案手法による学習効果について示す。

2. 関連研究

ABN は、推論時の判断根拠となるアテンションマップを特徴マップに乗算することで認識性能の向上を実現している。ABN はアテンションマップを介して視覚的な説明が可能であるが、学習データが少なかったり偏りがあると、注視領域が認識対象以外に発生する場合や認識対象を注視しない場合がある。三津原らは学習済みの ABN の誤ったアテンションマップを人の知見に基づいて修正し、教師データとして再学習することで正確なアテンションマップの獲得を実現した [2]。詳細画像識別タスクにおいて、人の知見を利用してアテンションマップを修正することで、物体のより詳細な領域に着目したアテンションマップの生成を実現し、優れた分類性能を達成した。

3. 提案手法

本研究は、エキスパートの知見に基づいてファインチューニングした ABN を教師とし、学習者がその判断根拠をインタラクティブに学ぶ学習法として、アテンションマップを編集し推論する教育アプリを作成する。

3.1 アテンション機構によるアテンションマップの修正

判断根拠をインタラクティブに学ぶために、ABN のアテンションマップを学習者が繰り返し修正する。その際、出力されるスコアの変動よりどの領域に着目すべきかという知識を学習者が自ら獲得する。学習者が編集したアテンションマップ $M_n \in \{0, 1\}$ を用いて、各クラスに対するスコアを出力するために、Feature Extractor から得られた特徴マップ $g(x)$ に対して重みづけを行う。式 (1) より重みづけされた特徴マップ $g'(x)$ を得る。

$$g'(x) = [g(x) \cdot M_n(x)] + g(x) \quad (1)$$

また、 $M_n \in \{0, 1\}$ は、学習者が注視した領域を 1、注視していない領域を 0 としている。 $g'(x)$ を用いて Perception Branch により、教育アプリで学習者に表示する各クラスにおけるスコアを出力する。学習者はこのスコアの変動から注視領域について学ぶ。

3.2 “AI から” 学ぶ教育アプリ

本アプリでは、学習者が編集したアテンションマップに対し、エキスパートの知見を組み込んだ AI がスコアを提示することで、学習者と AI がインタラクティブに学習することを可能とする。図 1 に提案手法を導入した教育アプリとその操作を示す。提案手法を導入した教育アプリを用いたインタラクティブに学習する手順を以下に示す。

- step.1 データセットから無作為に選択した画像を 1 枚ずつ表示し、学習者がクラス分類を行う。
- step.2 表示した画像に対応するラベルから正答を確認する。
- step.3 学習者がラベルに合わせて、画像の特徴を強調するようにアテンションマップの編集を行う。
- step.4 編集したアテンションマップを用いて、AI が各クラスにおけるスコアを推論する。
- step.5 学習者がラベルに合わせたスコアを確認する。
- step.6 クラス分類をしたスコアの変化より、自身の編集したアテンションマップを時系列的に比較し、スコアが向上する注視領域を学習する。

step.3 から step.6 の処理を繰り返すことで、学習者が AI によりインタラクティブに学習する。また、学習者がラベルに対して正答した場合は step.3 から step.6 を行わない。図 2 に、学習者が教育アプリによりインタラクティブに学習する様子を示す。図 2 より、適切な領域を追加することでスコアが上昇し、不適切な領域を追加するとスコアが下がることがわかる。このように能動的に学習することで、効率良く教育効果を得られると期待できる。

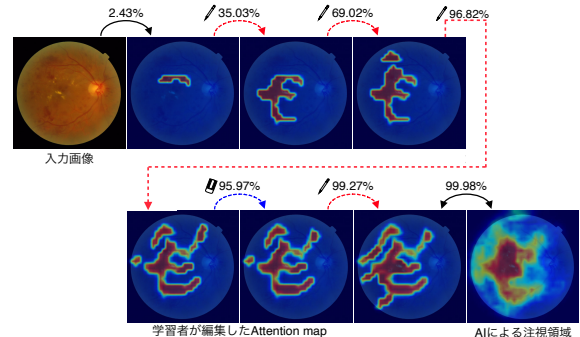


図 2：インタラクティブ学習によるスコアの変化

4. 評価実験

提案手法の効果を検証する。本研究では、眼底画像の疾患である糖尿病網膜症を対象とする。被験者 36 人を教育アプリなし、エキスパートの知見がない教育アプリ、提案手法による教育アプリの 3 つのグループに分け比較実験を行う。実験は、眼底疾患のデータセットから二値分類の出題を行い、正解率から学習効果を評価する。被験者が行う学習前と学習後のテストでは、Messidor のデータセットから正常な眼底画像を 30 枚、疾患のある眼底画像を 30 枚用いる。被験者が行う学習には、Indian Diabetic Retinopathy Image Dataset (IDRiD) のデータセットを使用する。正常な眼底画像を 124 枚、疾患のある眼底画像を 81 枚学習教材として用いて、被験者は学習を行う。

4.1 正解率による評価

3 つのグループの結果をそれぞれ表 1 に示す。ここで、高得点を 80 点以上と設定し、高得点者数を用いて分析を行う。

表 1 より、その他の学習法と比較し、提案手法を用いて学習することで、高得点者数が多いことが確認できる。また、アプリを使用しないグループとエキスパートの知見がない AI モデルを用いたグループでは、得点の標準偏差が

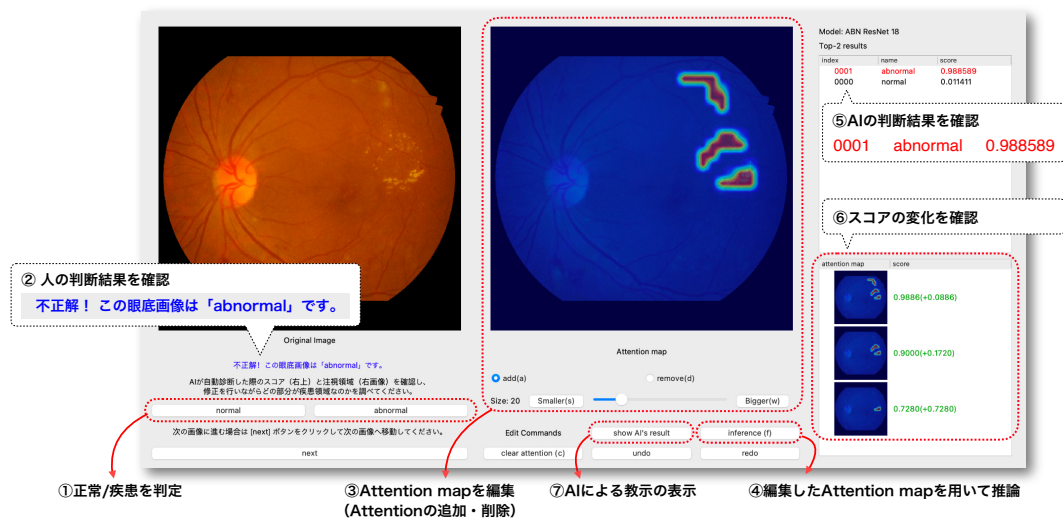


図 1：提案手法による教育アプリ

表 1：評価実験結果

	アプリなし		提案手法			
	学習前	学習後	学習前	学習後	学習前	学習後
80 点未満 [人]	12	9	12	11	12	5
80 点以上 [人]	0	3	0	1	0	7
平均 [%]	55.69	77.78	57.50	69.86	53.47	79.44
標準偏差	11.86	3.65	12.82	9.20	9.88	2.50

大きい。これは、学習やその後の解答が個人の能力に依存しているためと考えられる。一方、提案手法の教育アプリのグループは標準偏差が小さいことから、多くの学習者にとって教育効果が高い。

4.2 アテンションマップによる評価

被験者が編集したアテンションマップを比較し、適切な領域に注目しているかの学習効果を検証する。図 3 に各グループの被験者が編集したアテンションマップの例を示す。

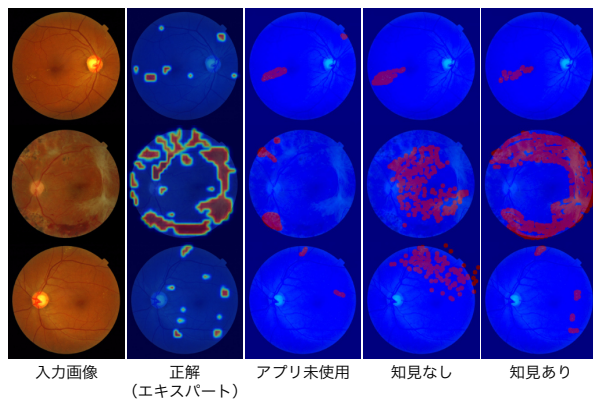


図 3：アテンションマップの比較図

提案手法により学習を行った被験者はエキスパートのアテンションマップと類似したアテンションマップを編集している。一方で、そのほかの手法では、異なる領域にアテンションを追加する場合や、一部エキスパートのアテンションと同じ箇所を編集しているものの、アテンションを追加していない領域などが見受けられる。

また、アテンションマップを用いた定量的評価として、30 枚の疾患画像に対するエキスパートと被験者のアテンションマップの疾患領域に対するクラス IoU を表 2 に示す。提案手法は学習後のクラス IoU が最も大きく、最もエキスパートの知見に近いアテンションマップを編集していることがわかる。以上より、エキスパートの知見を導入した AI モデルによる学習により被験者がインタラクティブにどこに注視すべき領域であるかを学習したことが確認できる。また、AI が示すアテンションマップにより、疾患判定の判断根拠を学習者に示すことで、高い学習効果が得られることを確認した。

表 2：エキスパートと被験者の IoU

	教育アプリなし		提案手法			
	学習前	学習後	学習前	学習後	学習前	学習後
平均	0.0958	0.1079	0.0895	0.0835	0.1107	0.1314

4.3 被験者からのフィードバック

教育アプリに対する定性的評価として、知見を導入した教育アプリを使用した被験者からの実験終了後のフィードバックを以下に示す。

- 能動的な学習により、疾患がどういったものかわかるようになった

一方、エキスパートの知見がない AI モデルを用いて学習した被験者からのフィードバックを以下に示す。

- 細かい領域に着目することができず、疾患の特徴に対する思い込みを修正するのが難しかった

以上より、教育アプリに知見を導入した AI モデルを併せて活用することで、判断根拠となる疾患領域を明確にし、提案手法の学習効果を促進することを示した。

5. おわりに

本研究では、エキスパートの知見を導入した深層学習モデルを教師として、学習者である人がその判断根拠を“AI から”学ぶ手法を提案した。提案手法では、ABN のアテンションマップ及びアテンション機構を活用し、ネットワークへエキスパートの知見を導入した。知見を導入した ABN を用いて、学習者が編集したアテンションマップに対しアテンション機構の活用による分類結果を確認することで、AI の判断根拠のもとに学習を行うことを可能とした。また、この学習法を容易に行うために、教育アプリケーションを作成し、被験者実験を行った。評価実験により、提案手法を用いることで、より効率的に学習を行うことができ、エキスパートの知見と同様の疾患領域に着目し判定することが可能となることを示した。本手法は、医療画像の疾患判定だけでなく、外観検査をはじめとする高い専門性を必要とするタスクへの活用が期待できる。

参考文献

- [1] H. Fukui, *et al.*, “Attention Branch Network : Learning of Attention Mechanism for Visual Explanation”, CVPR, 2019.
- [2] M. Mitsuhashi, *et al.*, “Embedding Human Knowledge into Deep Neural Network via Attention Map”, International Conference on Computer Vision Theory and Applications, pp. 626-636, 2021.

研究業績

- [1] 服部幸平, 平川翼, 山下隆義, 藤吉弘巨, “AI から学ぶ：エキスパートの知見を導入した DNN モデルを教師とする学習者のためのインタラクティブ学習法の提案”, 電子情報通信学会パターン認識・メディア理解研究会 (PRMU), 2022.