

エキスパートの知見を取り入れた マルチスケール・アテンション機構による疾患識別

坂下祐輔^{†‡} 平川翼[‡] 山下降義[‡] 藤吉弘亘[‡]

[†] 株式会社ニデック [‡] 中部大学

E-mail: yusuke_sakashita@nidek.co.jp

1 はじめに

畳み込みニューラルネットワーク (CNN) [1] では、推論時の注視領域をヒートマップで表現したアテンションマップとして獲得し、判断根拠の視覚的説明に用いている。視覚的説明の代表的な手法として、畳み込み層の応答値を用いてアテンションマップを獲得する Class Activation Mapping (CAM) [2] や、アテンションマップ獲得時に特定クラスの正值の勾配を重み値として用いる Gradient Weighted-CAM (Grad-CAM) [3]、視覚的説明のアテンションマップをアテンション機構へ応用する Attention Branch Network (ABN) [4] が提案されている。これらの視覚的説明の手法により、CNN の判断根拠の解釈が可能となってきた。特に、視覚的説明は医療における診断時の判断根拠として必要不可欠な情報になりつつある。しかし、Ground truth (GT) とこの視覚的説明である注視領域が一致しない場合が発生する。例えば、医療画像診断の場合、病気か否かを画像から判断する際は疾患領域に注目する。そのため、疾患領域とモデルが出力する注視領域の不一致は診断根拠を説明するのに深刻な問題を引き起こし、医療画像診断システムにおける信頼性の低下に繋がる。

この問題を解決するために、我々は視覚的説明とアテンション機構を導入した ABN の構造に着目する。ABN は、アテンション機構で取得されたアテンションマップを認識機構へ応用することで、認識精度の向上を目的としたネットワークである。さらに、アテンションマップを人手により変更することで、その変更したアテンションマップに対応する認識結果を得ることができる特徴を持つ (図 1)。このような ABN の特徴を用い、我々はモデルに対して人の知見を導入した再学習法を提案し、医用画像による評価を行った [5]。疾患を有する画像に対し、注視すべき疾患領域をアテンションマップとして与えることで、疾患の識別精度と注視領域改善により有効性を確認した。

しかし、一般的な識別モデルによる疾患識別では広範囲にわたる疾患領域は捉えることができるものの、図 2 のような微小な疾患を捉えることは困難である。これ

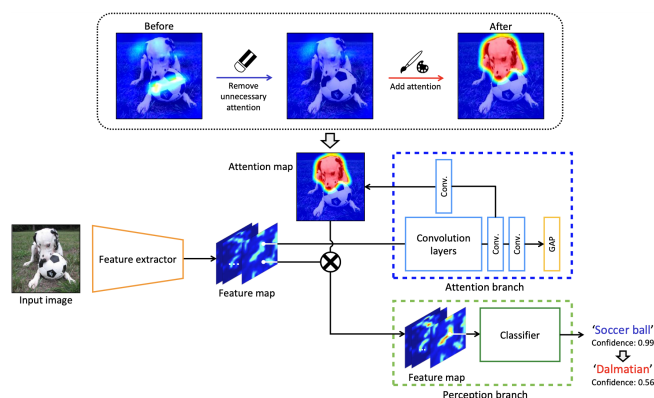


図 1 変更したアテンションマップを入力した ABN の識別結果

表 1 糖尿病網膜症 国際重症度分類		
重症度	眼底初見	
R0	網膜症なし	異常所見なし
R1	軽度非増殖網膜症	毛細血管瘤のみ
R2	中度非増殖網膜症	毛細血管瘤以上の病変を認める 重度非増殖網膜症より軽度
R3	重度非増殖網膜症	複数の網膜内出血、白斑、血管異常のいずれかを認める 増殖網膜症の初見を認めない
R4	増殖網膜症	新生血管、硝子体・網膜前出血のいずれかを認める

は、一般的な畳み込み層を複数持つモデルでは特徴マップのサイズが小さくなることにより、微小な領域を捉えることができないことが理由である。

ここで、眼底疾患である糖尿病網膜症 (DR: Diabetic Retinopathy) を例にあげる。糖尿病網膜症 (以降 DR) は微小動脈瘤、出血、白斑等の種類や量、形状に応じて表 1 に示すように 5 つのグレードに分類される [6]。しかし、DR のグレーディングは単なる疾患の識別問題ではなく、診断するために必要な病的異常は形状や外観、さらには疾患のサイズにかなりのばらつきがあるため、グレーディングを決定づけるための注目領域も疾患の種類により異なる。よって、動脈瘤や小領域の出血などの微小な疾患を捉えることができない従来の手法では、DR のグレード分類は困難である。

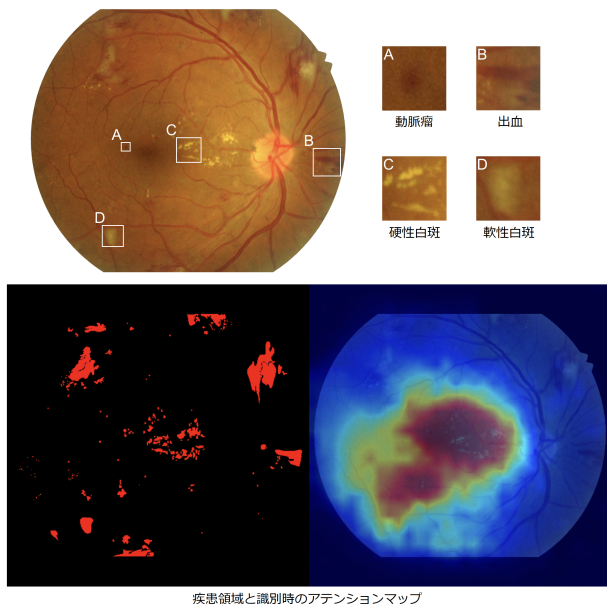


図2 眼底疾患の例と識別時の注視領域

そこで、本研究では眼底疾患であるDRのグレーディングを対象に提案手法の有効性を示す。提案手法は、ABNのアテンション機構の特徴を用いて医師、すなわちエキスパートの知見を導入し、DRグレーディング時にグレーディングラベルと同時に疾患領域も含めて学習する方法を用いる。また、DRのグレーディングに必要な微小な疾患の検出にも対応できるよう、マルチスケールのアテンション機構を取り入れた手法を提案する。提案する深層学習モデルでは、エキスパートの知見を基に人手で修正したアテンションマップとABNが出力した各層でのアテンションマップから学習誤差を算出し、ABN各層の識別に対応するブランチをファインチューニングする。エキスパートの知見とマルチスケールの機構を取り入れたアテンションマップを学習することで、より診断における説明性の高いアテンションマップの出力と識別性能の向上が期待できる。さらに、不適切な学習サンプルに対する修正コストを下げることを目的として、アクティブラーニングの観点から効率の良い学習サンプル選定に対する評価を行う。

以下に、本稿の貢献を示す。

- マルチスケール・アテンション機構に対しエキスパートの知見を与えてファインチューニングすることにより、各層で最適な疾患のサイズに応じたアテンションマップの獲得と、微小疾患の識別への対応
- アクティブラーニングの観点から効率の良い学習サンプルを選定することで、少ない修正コストでの識別精度の向上と疾患領域への注視の改善

2 提案手法

本研究は、エキスパートの知見により修正したアテンションマップを用いてABNをファインチューニングする。ファインチューニングによりエキスパートの知見を取り入れたアテンションマップを学習することで、ABNがエキスパートの知見に近いアテンションマップを出力するモデルになる。また、アテンション機構をマルチスケールにすることで、微小な疾患領域を考慮した疾患識別が可能になる。本章にて提案手法の各処理について詳細を述べる。

2.1 マルチスケール・アテンション機構

DRを識別するためには、形状や外観のばらつきが大きい疾患を捉える必要がある。しかし、畳み込み層を複数有するCNNでは、最終的な特徴マップのサイズは非常に小さくなる。そのため、微小な疾患領域を捉えることは困難であり、DRグレーディングには適さないという問題がある。そこで、本手法は大小様々なサイズの疾患に対応するため、マルチスケールのアテンション機構を用いる。図3にResNetをベースにしたマルチスケール・アテンション機構を用いたモデルの構成を示す。モデルは、3つのブロックで構成されたFeature Extractorと、それぞれのブロックに対応するAttention branch, Perception branchから成る。各Feature Extractorから出力された特徴マップをAttention branchに入力することで、各層に応じたサイズのアテンションマップを取得する。続いて、取得したアテンションマップとFeature Extractorの特徴マップを統合しPerception branchに入力することで各層に応じた特徴量が得られる。最後に、各Perception branchから得られる特徴量を統合することで、最終的なDRのグレーディングラベルを得る。

2.2 ファインチューニングに用いるデータの選定とアテンションマップの修正

図4に提案手法によりアテンションマップを修正してファインチューニングするまでの流れを示す。はじめに、学習サンプルと教師ラベルを基にABNを学習し、学習済みのABNを用いてアテンションマップを収集する。ここで、以下に該当するアテンションマップを収集する。

- 誤識別した学習サンプルに対応するアテンションマップ
- 識別は正しいがアテンションマップが疾患領域を捉えていない学習サンプルに対応するアテンションマップ

次に、収集したアテンションマップをエキスパートの知見を基にし、疾患領域を注視するようなアテンションマップへと修正する。しかし、不適切なデータに対し

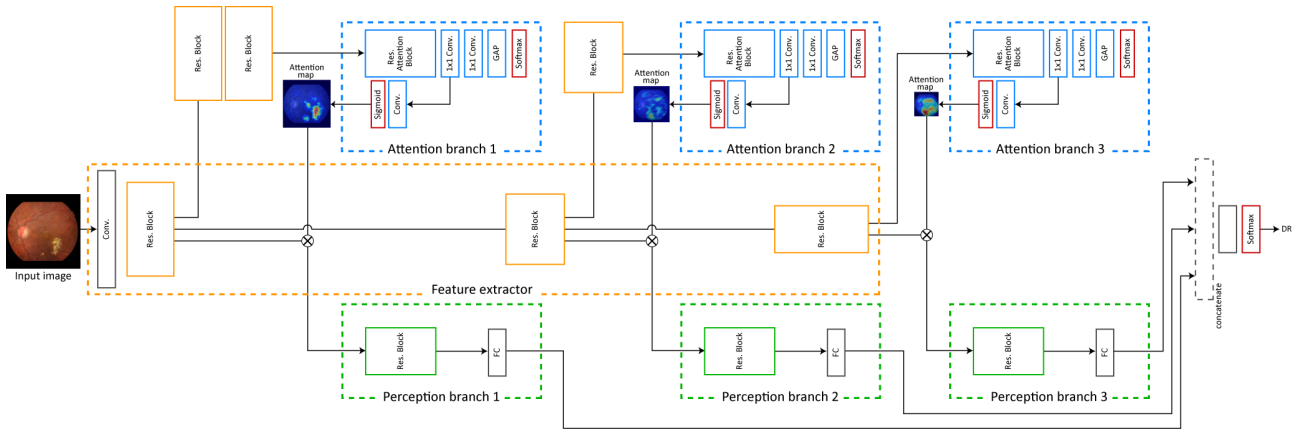


図3 マルチスケール・アテンション機構を導入したモデル

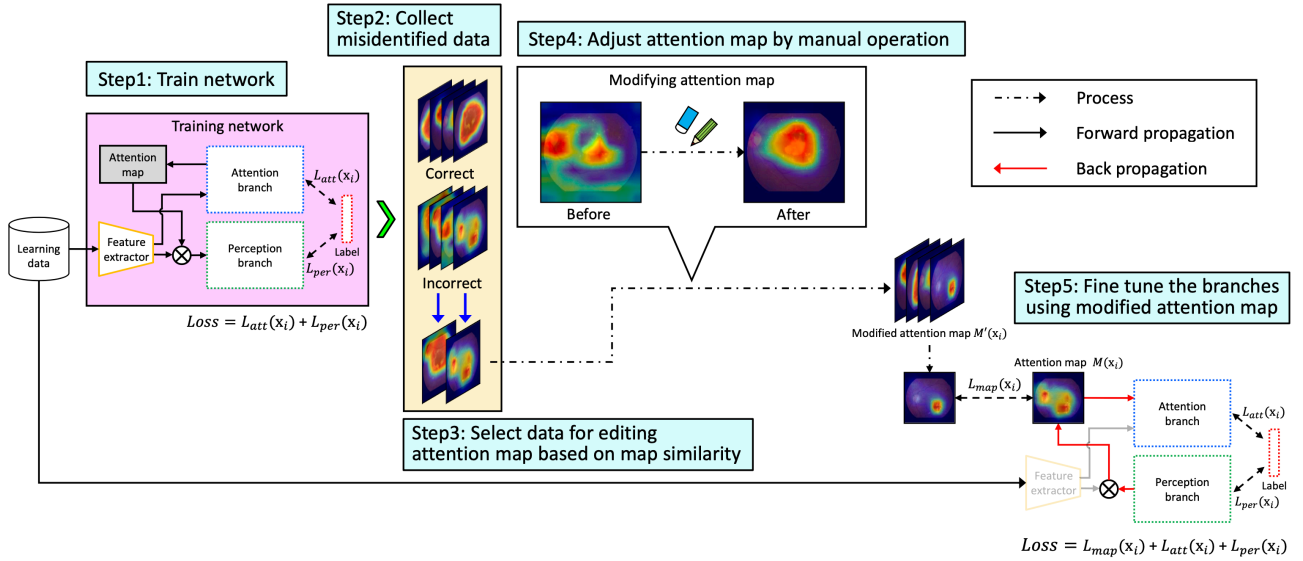


図4 アテンションマップの修正とファインチューニングの流れ

て全てのアテンションマップを修正してファインチューニングするのは効率的ではない。そこで、アクティブ・ラーニングの観点から修正すべき学習サンプルを選定を行う。その後、選定したデータに対する修正したアテンションマップを用いて ABN の Attention branch と Perception branch をファインチューニングする。ファインチューニングする際、ABN の Attention branch と Perception branch の学習誤差に加えて、出力されたアテンションマップと修正したアテンションマップを用いて学習誤差を算出する。

2.3 モデルのファインチューニング

エキスパートの知見を取り入れたアテンションマップを作成した後、これらのアテンションマップを用いて ABN をファインチューニングする。ファインチューニングでは、式 (1) のように従来の ABN の Attention branch と Perception branch の 2 つの学習誤差 L_{att} , L_{per} に加えて、出力されたアテンションマップとエキスパートの知見を組み込んだアテンションマップの学習誤差 L_{map} を追加する。

$$L_{all}(x_i) = \sum_{l=layer} (L_{att}(x_i)_l + L_{per}(x_i)_l + L_{map}(x_i)_l) \quad (1)$$

さらに、これらの学習誤差を各層の branch からの学習誤差として追加する。各層における L_{map} の導入により、ABN から出力されるアテンションマップはエキスパートの知見に近くなる。そのため、ABN はエキスパートの知見を考慮し、かつ疾患領域に適したアテンションマップを出力できるようにファインチューニングされる。

アテンションマップの学習誤差 L_{map} は、式 (2) のように L2 ノルム誤差を用いる。

$$L_{map}(x_i)_l = \lambda_l ||M'(x_i) - M(x_i)_l||_2 \quad (2)$$

ここで、 $M(x_i)$ は出力されたアテンションマップ、 $M'(x_i)$ はエキスパートの知見を取り入れたアテンションマップを示している。 $M(x_i)$ と $M'(x_i)$ の 2 つのアテンションマップの要素毎に誤差を求めることで、エキスパートの知見に近いアテンションマップを出力するように Attention branch が学習される。ここで、 λ は誤差関数 $L_{map}(x_i)$ を調整する係数である。

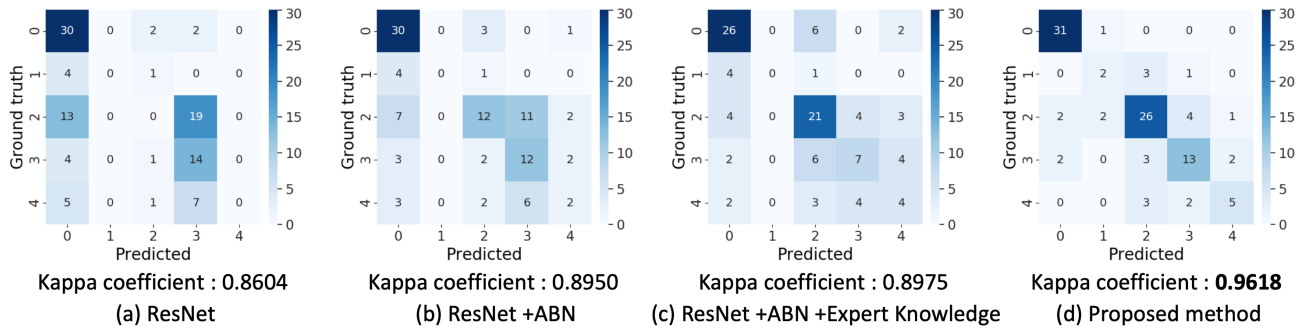


図 5 各手法における DR グレーディング結果

ABN をファインチューニングする際に、提案手法では各層のブランチを対象に学習する。入力画像から特徴マップを抽出する Feature extractor は、ファインチューニング時にパラメータを更新しない。

3 評価実験

評価実験では、Indian Diabetic Retinopathy Image Dataset (IDRiD) [7] を用いて、DR グレーディングにより提案手法の有効性を検証する。IDRiD は糖尿病網膜症における疾患のセグメンテーションやグレーディングを対象とした眼底画像データセットである。IDRiD を用いて限られたデータからでも識別精度が向上することに加え、疾患識別に必要な正しい疾患領域に注視することを示す。注視領域の検証として、アテンションマップを数値的に評価するために Deletion metric, Insertion metric, 修正したアテンションマップとモデルが出力したアテンションマップとの類似度を用いる。また、全ての学習サンプルにエキスパートの知見を反映させてファインチューニングを行うことは非現実的である。そのため、与えるべき学習サンプルを選定した効率の良いファインチューニングについても評価を行う。

3.1 糖尿病網膜症グレーディング評価

ここでは、IDRiD の DR グレーディングデータセットを用いて DR グレーディングの評価を行う。DR グレーディングデータセットは 413 枚の学習サンプルと 103 枚の評価サンプルからなり、全てのサンプル画像に対して 0 から 4 のグレードがラベル付けされている。これらの学習サンプルに対し、エキスパートの知見を基に正しいアテンションマップを生成し、2.2 節で述べたモデルのファインチューニングを行う。なお、評価に使用するモデルは ResNet18 をベースとし、ResNet18, ResNet18+ABN, ResNet18+ABN に修正したアテンションマップを与えてファインチューニングしたモデル (シングルスケール), マルチスケール・アテンション機構を導入した提案手法で比較を行う。

図 5 に各手法における DR グレーディングの結果、図 6 に提案モデルによる各 branch から出力されたアテ

ンションマップを示す。DR グレーディングの定量評価には、正解 (GT) と推測値の一致度を測る評価指標の一つである重み付き kappa 係数 [13] を用いた。提案手法では、第 3 ブランチで帯域に及ぶ領域を注視する一方、第 1, 第 2 ブランチでは細かな疾患領域に注視していることが分かり、各ブランチで DR グレーディングに必要な疾患を捉えていることが分かる。さらに、kappa 係数において、提案手法は正解の DR グレードとの一致度が非常に高いことが確認できた。

本実験によって、エキスパートの知見を反映することで DR グレーディング時に注視すべき領域を有効的に学習させ、さらにマルチスケール・アテンション機構によって DR グレーディングに必要な大小様々な形状の疾患の検出に有効であることを確認した。

3.2 注視領域の評価

ここでは、注視領域の評価としてアテンションマップを定量的に評価する。定量的評価には、Deletion metric, Insertion metric, 修正したアテンションマップとモデルが出力したアテンションマップとの類似度を用いる。なお、本検証には Feature extractor の最下層の Attention branch から得られたアテンションマップを用いる。

Deletion metric と Insertion metric は、Petsiuk ら [8] が提案する評価方法である。Deletion metric は、アテンションマップの注目度の高い領域を入力画像から徐々に削除することによるスコアの低下を測定する。したがって、スコアが低いほど説明性が高いことを意味している。一方で、Insertion metric は重要度の高い領域を徐々に追加することによるスコアの増加を測定する。したがって、スコアが高いほど説明性が高いことを意味している。修正したアテンションマップとモデルから出力されたアテンションマップとの類似度による評価では、平均二乗誤差により類似度を測定する。そのため、誤差が 0 に近いほど修正したアテンションマップとの類似度が高いことを示す。そこで、エキスパートである医師の知見によってラベル付けされた疾患領域のセグメンテーションラベルを基に、修正したアテンションマップとモデルが出力したアテンションマップ

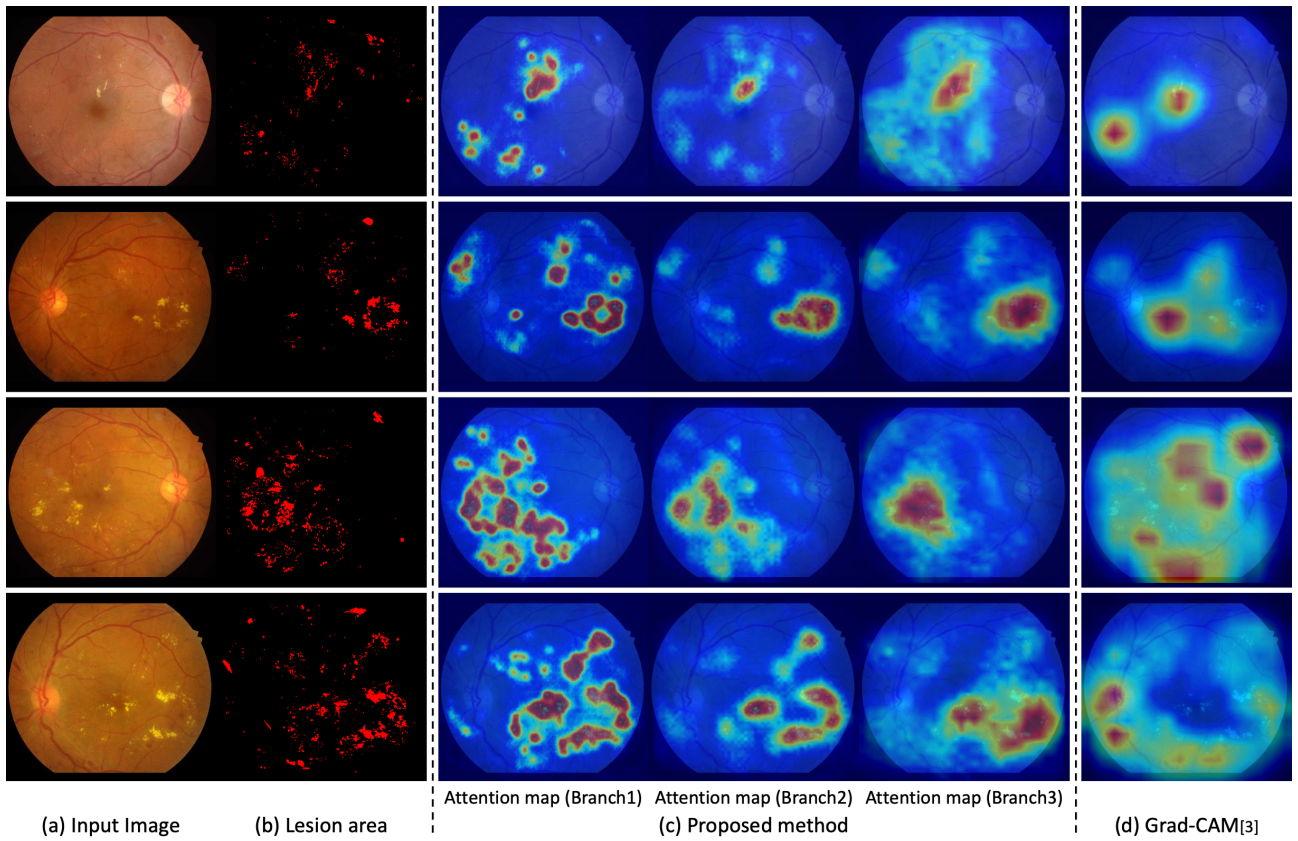


図 6 各ブランチにおけるアテンションマップ

表 2 各手法における Deletion スコアと Insertion スコアの比較

	AlexNet		ResNet18		ResNet34	
Method	Deletion	Insertion	Deletion	Insertion	Deletion	Insertion
Grad-CAM [3]	0.3058	0.6201	0.2958	0.7085	0.2854	0.7465
ABN [4]	0.4201	0.6028	0.3900	0.6878	0.3844	0.6524
Proposed	0.2732	0.8704	0.2978	0.8832	0.3102	0.8923

との類似度を測定する。

表 2 および表 3 に、それぞれ Deletion と Insertion のスコアとアテンションマップの類似度を示す。表 2 に示すように、提案手法は最も高い Insertion score を取得したことから、疾患識別に重要な領域を注視していることが確認できる。さらに、表 3 に示すように、提案手法はエキスパートによる疾患領域のセグメンテーションに最も適切に注視しているアテンションマップを出力していることが分かる。この結果から、提案手法は限られたデータにおいても、疾患識別に対する注視領域の改善に有効であることが確認できる。

3.3 学習サンプルの選定評価

実際に提案手法によりファインチューニングさせる場合、誤識別および不適切なアテンションマップに対する全ての学習サンプルのアテンションマップを修正して与えるのは非現実的である。そこで本実験では、以下の 2 つの方法によって与えるべき学習サンプルを選定した効率の良いファインチューニングの条件を評価する。

- 学習サンプルを confidence の低い順にソーティングし、confidence の低いサンプルからファインチューニング
- 学習サンプルのアテンションマップと Ground truth のマップの二乗誤差を高い順にソーティングし、類似度の低いサンプルからファインチューニング

その後、アクティブラーニング [9, 10, 11, 12] の観点から学習結果に応じて修正する学習サンプル数を一定の割合で増やし、学習サンプル数と識別率、アテンションマップの類似度の関係を検証する。

Fig. 7 に修正したアテンションマップをモデルに与えた学習サンプル数の割合と識別率、アテンションマップの類似度を示す。本実験で用いたデータセットにおいて、識別率の向上に寄与する選定条件は confidence を基に選定、アテンションマップの二乗誤差を基に選定共に、全学習サンプルのうち 30% である。それに対し、注視領域の向上に寄与する選定条件は confidence を基に選定した場合は全学習サンプルの 70% であるのに対し

表3 各手法におけるアテンションマップの二乗誤差

Method	AlexNet	ResNet18	ResNet34
Grad-CAM [3]	0.1873	0.1521	0.1329
ABN [4]	0.1193	0.1241	0.1309
Proposed	0.0873	0.0893	0.0927

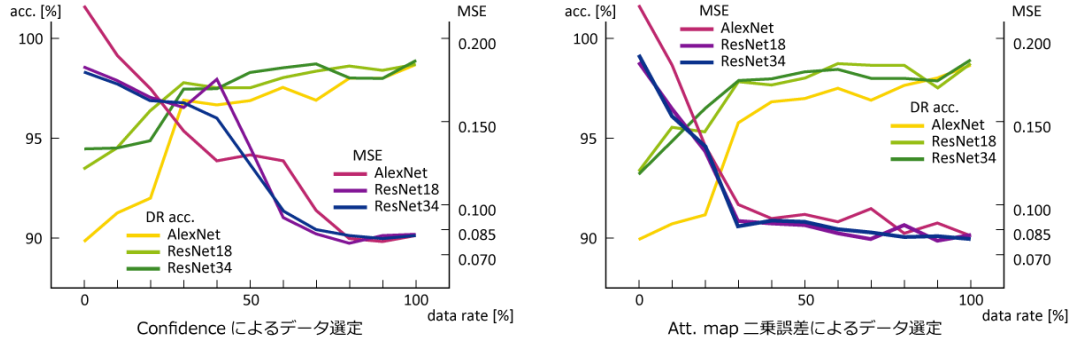


図7 学習サンプルの割合に対する識別率とアテンションマップの二乗誤差

し、アテンションマップの二乗誤差を基に選定した場合は30%から向上がみられた。したがって、注視領域を含めた疾患識別率の向上に寄与する学習サンプルの選定条件は、アテンションマップを基準に選定することが有効であることが確認できる。

4 おわりに

本稿では、エキスパートの知見を基に手動で修正したアテンションマップを出力するようにモデルをファインチューニングし、さらに異なるサイズの疾患に対応するためにマルチスケールのアテンション機構を導入したモデルを提案した。評価実験により、エキスパートの知見によって修正すべき学習サンプルをアテンションマップの類似度を基に30%程度を選定してファインチューニングさせることで、異なる形状の疾患を捉えることが必要なDRグレーディングにおいて、グレーディング精度が向上したことを示した。また、アテンションマップの定量評価によって、診断時にエキスパートが注視する疾患領域と同等のアテンションマップを出力するように改善できたことを確認した。これは、診断根拠における説明性の向上にもつながる。

本手法は、エキスパートの知見を有効的に利用しモデルに組み込むことで、医療等の重要な判断や明確な説明が求められる分野での活躍が期待できる。さらに、人の知見を取り入れた効率の良い学習が可能なことから、医療分野における希少疾患の識別や製造現場における異常検知等、データ取得が困難な分野への応用へも期待できる。

参考文献

[1] K. Alex, I. Sutskever, et al. : Imagenet classification with deep convolutional neural networks. In Neural

Information Processing Systems (NIPS), 2012.
 [2] B. Zhou, A. Khosla, et al. : Learning deep features for discriminative localization. Computer Vision and Pattern Recognition (CVPR), 2016.
 [3] S. Ramprasaath R., C. Michael, et al. : Grad-cam: Visual explanations from deep networks via gradient-based localization. In International Conference on Computer Vision (ICCV), 2017.
 [4] H. Fukui, T. Hirakawa, et al. : Attention branch network: Learning of attention mechanism for visual explanation. Computer Vision and Pattern Recognition (CVPR), 2019.
 [5] 三津原, 福井, 坂下, 他: Attention map を介した Deep Neural Network への人の知見の組み込み. 画像の理解・認識シンポジウム (MIRU), 2019.
 [6] American Academy of Ophthalmology : Diabetic retinopathy preferred practice pattern. 2019.
 [7] P. Porwal, S. Pachade, et al. : Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. 2018.
 [8] V. Petsiuk, A. Das, and K. Saenko. Rise : Randomized input sampling for explanation of black-box models. In Proceedings of the British Machine Vision Conference (BMVC), 2018.
 [9] D. Angluin. : Queries and concept learning. Machine learning, 2(4), pp.319–342, 1988.
 [10] L. E. Atlas, D. A. Cohn, and R. E. Ladner. : Training connectionist networks with queries and selective sampling. In Advances in neural information processing systems, pp.566–573, 1990.
 [11] M. Wang, and X. S. Hua. : Active learning in multimedia annotation and retrieval: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2(2), pp. 10–31, 2011.
 [12] K. Ksenia, S. Raphael, et al. : Learning Active Learning from Data. Advances in Neural Information Processing Systems (NIPS), 2017.
 [13] Fleiss J. L. and J. Cohen. : The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability. Educational and Psychological Measurement 33, pp. 613–619, (1973).