Grad-CAM に基づく視線一貫性指標の定義と深層学習モデルへの適用

米村 慶太†
† 有明工業高等専門学校 創造工学科 情報システムコース
Gauthier Lovic 研究室

Definition of a Grad-CAM-Based Gaze Consistency Metric and Its Application to Deep Learning Models

Keita Yonemura†
† Information System Course, Department of Creative Engineering, National Institute of Technology (KOSEN), Ariake College
Gauthier Lovic Laboratory

Abstract– 深層学習モデルは高い性能を示す一方で, その判断根拠が不透明であることが課題とされており，医用画像解析をはじめとする高信頼性が要求される分野では，説明可能人工知能（Explainable AI: XAI）の重要性が高まっている．中でも Grad-CAM は，分類に寄与した画像領域を可視化する代表的手法であるが，その可視化結果の妥当性や安定性を定量的に評価する指標は十分に確立されていない.
本研究では，Grad-CAM によって得られる可視化結果をモデルの「視線」と捉え，同一クラスに属する入力画像間でその視線がどの程度一貫しているかを評価する指標として，視線一貫性（gaze consistency）を定義する．提案指標は，正解予測されたサンプルに対する Grad-CAM マップ間の距離に基づいて算出され，モデル性能との関係を解析的に評価可能である．さらに，本指標を正則化項として学習過程に組み込むことで，視線の一貫性を直接制御する手法を提案する.
複数のニューラルネットワークモデルおよび医用画像データセットを用いた実験の結果，分類性能と視線一貫性の関係はデータセット特性に強く依存し，一貫性の向上が必ずしも性能改善につながらないことが明らかとなった．一方で，提案する正則化は特定条件下において視線分布を安定化させる効果を示した．これらの結果は，Grad-CAM 可視化を評価・制御の対象として扱うことの有効性と同時に，XAI 指標の解釈における慎重さの必要性を示唆している.

Keywords: 画像分類, 説明可能 AI（XAI）, Grad-CAM, 視線一貫性, モデル解釈, 医用画像解析, CNN

目次

1 緒論

近年, 深層学習, 特に畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）の発展により，画像分類をはじめとする多くの認識タスクにおいて，人間の性能を上回る精度が達成されている．これらの手法は, 医療画像診断, 自動運転, 監視システムなど, 高い信頼性が要求される分野への応用が進んでおり，実社会における重要性は年々増している．

一方で，深層学習モデルは高い表現能力を有する反面, その内部動作が複雑であるため, モデルがどのような根拠に基づいて予測を行っているのかを人間が理解することは容易ではない．このようなブラックボックス性は, 誤判定時の原因分析や, 利用者からの信頼性確保，さらには説明責任が求められる場面において大きな課題となっている．

この問題に対し，近年ではモデルの判断根拠を可視化する手法が数多く提案されている．その代表的な手法として Grad-CAM が挙げられ, 予測に寄与した画像領域をヒートマップとして可視化することが可能である. これにより, モデルが注目している領域を直感的に理解できるようになった．

しかし，Grad-CAM による可視化結果は定性的に解釈されることが多く, その「良さ」を客観的に評価する指標について確立されているとは言い難い. このため, 可視化結果が信頼できる根拠となっているのか，あるいは単なる見かけ上の説明に留まっているのかを判断することは容易ではない．

そこで本研究では，Grad-CAM によって得られる注目領域を「モデルの視線」と捉え, 構図の類似した画像群を対象として，視線の空間的一貫性を定量的に評価する手法を検討する. さらに, 性能の異なる複数の分類モデルを比較することで，分類性能と視線一貫性との関係を分析する．

加えて，視線の一貫性を損失関数として学習過程に組み込むことで，視線を明示的に制御した場合に，分類性能およびモデルの挙動がどのように変化するかを検証する. 本研究は, 視線制約による性能向上を前提とするものではなく，性能低下やトレードオフが生じる場合も含めて，その影響を実験的に明らかにすることを目的とする．

本研究を通じて, 深層学習モデルが「正しい予測を行うこと」と「妥当な根拠に基づいて予測を行うこと」との関係について知見を提供し, 説明可能性と性能評価の橋渡しを行うことを目指す．

本研究において筆者は，Grad-CAM を基盤技術として用いながらも

1. クラス内の可視化結果の一貫性を定量化する新たな指標の設計
2. 当該指標を用いた可視化結果の解析的評価手法の構築
3. 指標を正則化項として学習過程に組み込む手法の実装
4. 複数データセット・モデルに対する実験および結果の分析

を一貫して行った．

2 関連研究

2.1 深層学習による画像分類

画像分類は，深層学習の発展とともに著しい進歩を遂げてきた分野の一つである．特に畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）は, 局所的な特徴抽出と階層的表現学習により，従来手法を大きく上回る性能を示してきた．

その代表例として，ImageNet Large Scale Visual Recognition Challenge（ILSVRC）において飛躍的な性能向上を示した AlexNet [1] が挙げられる．その後，ネットワークを深層化した VGG [2]，残差結合を導入した ResNet [3] などが提案され, 画像分類精度は継続的に向上してきた. これらのモデルは, 医用画像解析や自動運転, 監視システムなど，さまざまな応用分野において実用化が進んでいる．

一方で, モデルの高性能化に伴い, その内部挙動や判断根拠はますます複雑化している. CNN は高次元かつ非線形な特徴表現を学習するため，なぜ特定の予測結果が得られたのかを人間が直感的に理解することは困難である. この「ブラックボックス性」は，特に医療，司法，自動運転分野など，高リスクな意思決定（High-stakes decisions）において AI への信頼，安全性，非差別性を確保するために重要な課題となっている．

このような背景を踏まえ，本研究では分類モデル自体の性能向上を主目的とはせず, 既存の CNN がどのような根拠に基づいて判断を行っているのか，すなわちモデルの判断挙動に着目する．

2.2 モデル解釈手法（Grad-CAM, CAM, LRP など）

深層学習モデルのブラックボックス性に対処するため, 近年では Explainable AI（XAI）に関する研究が活発に行われている [4]．XAI の目的は，モデルの予測結果を人間が理解可能な形で説明し，信頼性や妥当性を検証可能にすることである．

画像分類における代表的な解釈手法として，Class Activation Mapping（CAM）が提案されている [5]．CAM は，最終畳み込み層の特徴マップと分類層の重みを用いて，特定クラスに寄与した画像領域を可視化する手法である．ただし，CAM はネットワーク構造に制約を課すため，既存モデルへの適用が容易でないという制限がある．

この制約を緩和した手法として，Grad-CAM が提案された [6]．Grad-CAM は，クラススコアに対する勾配情報を用いて特徴マップを重み付けすることで, ネットワーク構造に依存せず, 事後的に可視化を行うことが可能である．この汎用性の高さから，Grad-CAM は現在最も広く利用されている可視化手法の一つである．

そのほかにも，Layer-wise Relevance Propagation（LRP）[7] や Integrated Gradients [8] など，多様な説明手法が提案されている．しかし，これらの多くは「なぜこの予測が得られたの

か」を説明することを主眼としており, 可視化結果そのものの性質や安定性を体系的に評価する枠組みは十分に確立されていない.

本研究では, 汎用性が高く, 多くの既存研究で採用されている Grad-CAM を対象とし, その可視化結果の性質を定量的に分析する.

2.3 可視化結果の評価・一貫性・信頼性に関する研究

XAI 手法の普及に伴い，可視化結果の妥当性や信頼性を検証する重要性が指摘されるようになってきた. 可視化が直感的に「もっともらしく」見えることと，実際にモデルの判断根拠を正確に反映していることは必ずしも一致しない [9] .

この問題に対し，ランダム化テストや摂動解析を用いて説明手法の信頼性を検証する研究が報告されている. また, 同一入力に対する説明結果の安定性や，入力の微小変化に対するロバスト性に着目した研究も存在する [10] .

一方で,「同一クラスに属する複数の入力に対して, モデルが一貫した注目領域を形成しているか」という観点から説明結果を評価した研究は限定的である. 多くの場合, 可視化は個々の入力に対する事後的な説明に留まり，クラスレベルでの構造的な一貫性は十分に議論されてこなかった.

さらに, 可視化結果を定量的指標として定義し, それを学習過程に直接組み込む試みは少ない．説明可能性と学習性能の関係を体系的に分析した研究は，依然として発展途上にある.

本研究は，Grad-CAM に基づく注目領域の視線一貫性に着目し,

- その一貫性を定量的に評価する指標を定義し
- 分類性能との関係を解析的に評価するとともに
- 学習過程に正則化項として組み込んだ場合の影響を検討する

ことを目的とする点で，既存研究と差別化される.

3 分析手法

深層学習モデルの判断根拠を可視化する手法として，Grad-CAM をはじめとするモデル解釈手法が広く用いられている. これらの手法は, モデルが入力画像のどの領域に注目して予測を行っているかを示すものであり，主に分類結果の妥当性や信頼性を人間が確認するための補助的情報として位置づけられてきた.

従来の多くの研究では，モデル性能の最適化を主目的とし, その後に視線情報を計測･分析するアプローチが採用されている. このアプローチでは, まず分類精度の向上を目的としてモデルを学習し，学習後に Grad-CAM を適用することで，正解予測時にモデルが注目した領域を可視化する．得られた可視化結果は，モデルが妥当な特徴に基づいて判断しているかを確認するために利用されるが，視線情報自体は学習過程に直接的な影響を与えない．すなわち，Grad-CAM はあくまで事後的な分析手段として扱われている.

一方で, 本研究ではこれとは異なる立場として, 視線情報の性質を学習・評価の対象として明示的に扱うアプローチに着目する．具体的には，Grad-CAM によって得られる注目領域分布に着目し, その空間的な一貫性や安定性を定量的に評価可能な指標として定式化する. さらに, この視線一貫性を損失関数の一部として学習に組み込むことで，モデルの判断過程そのものに制約を与える.

このようなアプローチは,「視線の調整を通じてモデル性能を評価する」という観点に立つものであり，従来の「性能を調整した後に視線を観察する」枠組みとは方向性が異なる. 視線一貫性を考慮した学習を行うことで，単に正解ラベルを当てるだけでなく，クラスごとに一貫した判断根拠を持つモデルの獲得が期待される. また, 視線情報を定量的に扱うことで，分類性能と判断根拠の安定性との関係を体系的に分析することが可能となる.

本研究では，これら二つのアプローチを明確に区別した上で，後者の立場を採用し，Grad-CAM に基づく視線一貫性指標の定義およびそれを用いた学習手法を提案する．以降では，まず Grad-CAM の定義を整理し, 次に視線一貫性を定量化する指標を導入した上で，それを正則化項として組み込んだ学習手法について詳述する.

3.1 問題設定

本研究では，入力画像 $x \in \mathbb{R}^{H \times W \times C_{\text{in}}}$ に対して，クラスラベル $y \in C$ を予測する画像分類問題を対象とする．ここで，H は画像の高さ，W は画像の幅，C_{in} は入力チャネル数（例: RGB 画像の場合は 3）を表す．また，$C = \{1, ..., K\}$ は分類クラスの集合であり，K はクラス数を表す.

分類モデル f_θ は，パラメータ θ を持つ深層ニューラルネットワークであり，入力画像 x に対してクラス確率分布 $f_\theta(x) \in \mathbb{R}^K$ を出力する.

本研究の目的は, 分類精度そのものだけでなく, モデルが予測に至る過程において入力画像のどの空間領域に注目しているかという挙動を分析・制御することである．そのため，Grad-CAM によって得られるクラス判別に寄与する空間的注目領域を，モデルの「視線」と捉える.

本研究では，この「視線」の一貫性（consistency）に着目し，以下の二つの立場から検討を行う.

1. 学習済みモデルに対する解析的評価としての一貫性評価
2. モデル学習時に一貫性を誘導するための損失関数としての一貫性制約

これら二つは目的および要請される性質が異なるため，本研究では異なる定義に基づいてそれぞれを設計する. 特に, 後者の学習時一貫性については, ミニバッチ学習における安定性を考慮し, 評価用の定義をそのまま用いない構成を採用する.

3.2 ベースラインモデル

分類モデルとしては，畳み込みニューラルネットワーク（CNN）をベースラインとして用いる．本研究では，モデル構造そのものの新規性を主張するものではないため，既存の標準的な構成を採用する．

具体的なネットワーク構成やパラメータ設定については，4 章 で詳述する．

3.3 Grad-CAM の定義

Selvaraju らによって提案された Grad-CAM は，分類モデルの勾配情報を用いて，入力画像の各空間位置が特定のクラス予測にどの程度寄与したかを可視化する手法である [6] ．対象とする畳み込み層の特徴マップと，そのクラススコアに対する勾配を組み合わせることで，クラスごとの注目領域を生成する．

クラス $c \in C$ に対する Grad-CAM は，対象とする畳み込み層の特徴マップ $A^k (k = 1, ..., M)$ と，そのクラススコア y^c に対する勾配を用いて，次式で定義される．

$$\begin{aligned} \alpha_k^c &= \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \\ \mathrm{CAM}^{(c)} &= \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k \right) \end{aligned} \tag{1}$$

ここで，$A^k \in \mathbb{R}^{H' \times W'}$ は対象とする畳み込み層の k 番目の特徴マップを表し，$A_{i,j}^k$ はその空間位置 (i, j) における活性値である．添字 i, j はそれぞれ高さ方向および幅方向の空間インデックスとする．また，$y^{(c)}$ はクラス c に対応するモデルの出力スコア（ソフトマックス適用前）を表す．Z は特徴マップの空間サイズに基づく正規化定数であり[1]，M は対象層における特徴マップ数を表す．

本研究では，この $\mathrm{CAM}^{(c)}$ をモデルがクラス c の予測に際して注目した空間領域，すなわち「モデルの視線」として扱う．

本研究では，得られた Grad-CAM を空間的に正規化し，次式により $\widehat{\mathrm{CAM}}^{(c)} \in [0, 1]^{H' \times W'}$ を得る．

$$\widehat{\mathrm{CAM}}^{(c)} = \frac{\mathrm{CAM}^{(c)} - \min\left(\mathrm{CAM}^{(c)}\right)}{\max\left(\mathrm{CAM}^{(c)}\right) - \min\left(\mathrm{CAM}^{(c)}\right) + \varepsilon} \tag{2}$$

ここで，ε は，数値安定化のための小さな正の定数である．

本研究では，以降の議論において，特に断りのない限り，正規化後の Grad-CAM を CAM と表記する．

3.4 視線一貫性の定義および評価手法

3.4.1 分類性能指標

本研究では，提案手法の有効性を分類性能と視線挙動の両面から評価する．分類性能については，モデルが入力画像に対して正しいクラス予測を行えているかを評価する目的で，正答率（Accuracy）および Cohen's κ 係数を用いる．

正答率は最も基本的かつ直感的な指標であり，以下の式で定義される．

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[y_i = \hat{y}_i] \tag{3}$$

ここで，N は評価データ数，y_i は真のラベル，$\hat{y}_i$ はモデルによる予測ラベルを表す．また，$\mathbb{1}[...]$ は条件が真であるとき 1，偽であるとき 0 を返す指示関数である．

一方で，正答率はクラス数の増加やクラス分布の偏りに対して敏感であり，偶然による一致の影響を強く受ける可能性がある．特に，本研究のように複数のデータセットおよびモデル条件を比較する場合，偶然一致を考慮した指標を用いることが重要であり，正答率のみではモデルの本質的な識別能力を十分に評価できない．

そこで本研究では，主要な分類性能指標として Cohen's κ 係数 を採用する．κ 係数は，観測された一致率と偶然による一致率との差を正規化した指標であり，以下の式で定義される．

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

ここで，p_o は観測された一致率，p_e はクラス分布に基づいて算出される偶然一致率を表す． $\kappa \in [-1, 1]$ の範囲を取り，値が大きいほどモデルの予測が偶然を超えて真のラベルと一致していることを示す[2]．

本研究では，モデルの分類能力の比較・評価には κ 係数を主要指標として用い，正答率は補助的に併用する．

3.4.2 視線一貫性の基本定義

本研究では，分類結果そのものに加えて，モデルが予測に至る際に入力画像のどの空間領域に注目しているかに着目する．Grad-CAM により得られるクラスごとの注目分布を，モデルの「視線」と捉え，その一貫性 を定量的に評価・制御する枠組みを構築する．

データセット集合 $D = \{(x_i, y_i)\}_{i=1}^{|D|}$ に対し，クラス $c \in C$ に属するデータセット集合を

$$D_c := \{(x_i, y_i) \in D \mid y_i = c\} \tag{5}$$

と定義する．

[1] Z は一般的に $H'W'$ として定義される

[2] $\kappa = 0$ で偶然一致とみなされる

さらに，クラス c に属し，モデルが正解予測を行った入力画像群を

$$D_c^+ \coloneqq \{(x_i, y_i) \in D_c \mid \hat{y}_i = y_i\} \tag{6}$$

と定義する．

各入力画像 $x_i \in D_c^+$ に対し，クラス c に対する Grad-CAM によって得られる正規化済み注目分布を

$$\mathrm{CAM}^{(c)}(x_i) \in [0,1]^{H' \times W'} \tag{7}$$

と表す．

本研究における視線一貫性とは，**同一クラスに属する複数の入力画像に対して，モデルが空間的に類似した注目分布を示す性質**と定義する．

この概念に基づき，クラス c に対する代表注目分布を，次式により定義する．

$$\overline{\mathrm{CAM}}_D^{(c)} \coloneqq \frac{1}{|D_c^+|} \sum_{x_i \in D_c^+} \mathrm{CAM}^{(c)}(x_i) \tag{8}$$

さらに，個々の入力画像 x_i に対する注目分布と，そのクラスの代表分布との差を，視線偏差として次式で定義する．

$$\ell_{\mathrm{CAM}}(x_i, c) \coloneqq \frac{1}{H'W'} \left\| \mathrm{CAM}^{(c)}(x_i) - \overline{\mathrm{CAM}}_D^{(c)} \right\|_2^2 \tag{9}$$

ここで，ノルムは空間次元 $H' \times W'$ にわたる L2 ノルムであり，空間サイズ $H'W'$ により正規化されている．

3.4.3 視線一貫性の評価指標

前項で定義した視線一貫性の概念に基づき，ここでは モデル全体の挙動を解析的に評価する指標 として，データセット全体にわたる視線一貫性評価値を定義する．

クラスごとの正解サンプルの視線偏差を全ての画像について集約し，次式により最終的な視線一貫性評価値を定義する．

$$\mathcal{L}_{\mathrm{CAM}} \coloneqq \frac{1}{|D^+|} \sum_{c \in C} \sum_{x_i \in D_c^+} \ell_{\mathrm{CAM}}(x_i, c) \tag{10}$$

ここで，$D^+ = \{(x_i, y_i) \in D \mid \hat{y}_i = y_i\}$ は，モデルが正解予測を行った全入力画像の集合を表す．この値は，本研究における Grad-CAM の正規化定義の下ではおおよそ $\mathcal{L}_{\mathrm{CAM}} \in [0,1]$ の範囲を取り[3]，モデルはクラスごとに空間的に一貫した注目領域を用いて分類を行っているほど小さくなると解釈できる．

なお，本評価指標は，

- 正解ラベルが既知であること
- 各クラスに十分なサンプル数が存在すること

を前提としており，学習後のモデルに対する評価・比較を目的とした解析指標として用いる．

3.5 視線一貫性を考慮した学習手法

視線一貫性が分類性能やモデル挙動に与える影響を検証するため，本研究では視線一貫性を 正則化項として学習過程に組み込む実験も行う．

通常の分類損失を L_{cls} とすると，最終的な最適化対象は次式で与えられる．

$$\mathcal{L} \coloneqq L_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}} \tag{11}$$

ここで，λ は視線一貫性正則化の寄与度を制御するハイパーパラメータである．

ただし，前節で定義した評価指標はデータセット全体を前提としており，ミニバッチ学習に直接適用すると，

- クラス内サンプル数の不足
- バッチごとの不安定な推定

といった問題が生じる．

以上を踏まえ，学習時にはミニバッチ内で安定に計算可能な簡略化された一貫性損失 $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}}$ を次項で定義する．

λ 値に応じて，学習過程における視線一貫性の扱いは以下のように変化する．

- $\lambda > 0$ の場合:分類精度を維持しつつ，同一クラスに属する入力に対して類似した注目領域を持つように学習される．
- $\lambda = 0$ の場合:視線一貫性を考慮しない，通常の分類学習に対応する．
- $\lambda < 0$ の場合:視線一貫性損失を最大化する方向に学習が進み，同一クラス内で注目領域のばらつきが増大する．本研究では，この状態を便宜的に「視線の多様性」と呼ぶ．

なお，$\lambda < 0$ による視線の多様性は，視線一貫性の効果を検証するための対照条件として位置づけており，その性質に関する詳細な分析は本研究の主眼とはしない．

3.5.1 視線一貫性の損失指標

ミニバッチ学習において視線一貫性を損失関数として導入する場合，前節で定義した評価用指標 $\mathcal{L}_{\mathrm{CAM}}$ をそのまま用いることは適切ではない．

これを踏まえ，評価用一貫性指標と同一の概念に基づきつつ，ミニバッチ内で安定に計算可能な学習用視線一貫性損失 $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}}$ を以下のように定義する．

ミニバッチ集合 $B = \{(x_i, y_i)\}_{i=1}^{|B|}$ に対し，クラス $c \in C$ に属するデータセット集合を

$$B_c \coloneqq \{(x_i, y_i) \in B \mid y_i = c\} \tag{12}$$

と定義する．

さらに，クラス c に属し，モデルが正解予測を行った入力画像群を

[3] L2 ノルムの平均のため，上界が 1 であることが理論的に保証されるわけではない

$$B_c^+ := \{(x_i, y_i) \in B_c \mid \hat{y}_i = y_i\} \tag{13}$$

と定義する．

ここで，$|B_c^+| \geq 2$ を満たさないクラスについては，そのバッチにおける視線一貫性損失の計算から除外する．

クラス c に対する代表注目分布を，次式により定義する．

$$\overline{\mathrm{CAM}}_B^{(c)} := \frac{1}{|B_c^+|} \sum_{x_i \in B_c^+} \mathrm{CAM}^{(c)}(x_i) \tag{14}$$

さらに，個々の入力画像 x_i に対する視線偏差を次式で定義する．

$$\ell_{\mathrm{CAM}}^{\mathrm{train}}(x_i, c) := \frac{1}{H'W'} \left\| \mathrm{CAM}^{(c)}(x_i) - \overline{\mathrm{CAM}}_B^{(c)} \right\|_2^2 \tag{15}$$

クラスごとの正解サンプルの視線偏差をバッチ内の全ての画像について集約し，次式により最終的な学習用視線一貫性損失を定義する．

$$\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}} := \frac{1}{|B^{+'}|} \sum_{c \in C'} \sum_{x_i \in B_c^{+'}} \ell_{\mathrm{CAM}}^{\mathrm{train}}(x_i, c) \tag{16}$$

ここで，$C' = \{c \mid |B_c^+| \geq 2\}$ はバッチ内で視線一貫性損失の計算に用いるクラス集合を表し，$|B^{+'}| = \sum_{c \in C'} |B_c^+|$ はこれらのクラスに属する正解予測サンプル数の総和を表す．ここで，$B^{+'}$ は単に B_c^+ の和集合を表すものではなく，あくまで損失計算に用いる正解予測サンプル集合を示すものであることに注意されたい．

また，$|B^{+'}| = 0$，すなわち対象バッチ内に正解予測されたサンプルが存在しない場合には，式 (16) に基づく視線一貫性損失は定義できない．本研究ではこの場合の学習用視線一貫性損失を $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}} := 0$ と定義し，当該バッチにおいて本損失が学習に寄与しないように設計した．[4]

本定義には以下の特徴がある．

- クラスごとに独立して視線一貫性を評価するため，クラス間干渉が生じない
- クラス内平均との差のみを用いることで，データセット全体の統計量を必要としない
- ミニバッチサイズが小さい場合でも，条件を満たすクラスに対しては安定した勾配が得られる

このように，学習用視線一貫性損失は，評価用指標と同一の概念に基づきつつ，ミニバッチ学習に適合する形に再設計されたものである．

[4]本研究では $|B^{+'}| = 0$ の場合に損失を 0 と定義したが，代替案として，誤予測のみが生じた状況に対して一定のペナルティを与える目的で，定数値（例：1）を損失として与える設計も考えられる．しかし，本研究の視線一貫性損失は「正解予測に基づく説明の一貫性」を評価対象としているため，正解予測が存在しない場合には評価自体を無効化する設計がより自然であると判断した．

3.5.2 クラス平均視線の指数移動平均による安定化

前項で定義した学習用視線一貫性損失は，ミニバッチ内でのクラス代表注目分布 $\overline{\mathrm{CAM}}_B^{(c)}$ を基準として各サンプルの視線偏差を評価している．

しかし，ミニバッチサイズが小さい場合や，クラス数 K が大きい場合，クラスごとのサンプル数が偏っている場合には，$\overline{\mathrm{CAM}}_B^{(c)}$ の推定が不安定となり，学習が不安定になる可能性がある．

以上を踏まえ，各クラスの代表注目分布を指数移動平均（Exponential Moving Average: EMA）により保持し，ミニバッチ内の代表注目分布の代替または補助として用いる手法を採用する．

クラス c に対する EMA クラス CAM を $\widetilde{\mathrm{CAM}}^{(c)}$ とし，学習ステップ t において次式で更新する．

$$\widetilde{\mathrm{CAM}}_t^{(c)} = \alpha \widetilde{\mathrm{CAM}}_{t-1}^{(c)} + (1 - \alpha) \overline{\mathrm{CAM}}_{B,t}^{(c)} \tag{17}$$

ここで，$\alpha \in [0, 1]$ は 移動平均の平滑化係数であり，$\overline{\mathrm{CAM}}_{B,t}^{(c)}$ はステップ t におけるミニバッチ内のクラス c の代表注目分布を表す．

なお，当該ミニバッチにおいて $|B_c^+| < 2$ の場合には，当該クラスの EMA は更新されない．

学習用視線一貫性損失の計算においては，代表注目分布として $\widetilde{\mathrm{CAM}}^{(c)}$ を用いることで，

- クラス内視線の基準分布を時間的に平滑化できる
- ミニバッチごとのばらつきに起因する勾配ノイズを低減できる
- 極端に正解率が低い・クラスサンプル数が少ない状況でも安定した学習が可能になる

という利点が得られる．

この EMA に基づくクラス CAM は，評価指標として用いるものではなく，あくまで学習過程の安定化を目的とした補助的な手法である．

3.5.3 カリキュラム学習による視線一貫性制約の導入

視線一貫性損失を学習過程に導入する際，学習初期から強い制約を課すと，モデルが適切に特徴を学習できず，分類性能が著しく低下する可能性がある．

また，学習初期は正解率が低く，視線一貫性損失の計算に用いるサンプル数が不足しがちであるため，視線一貫性損失自体が不安定になるリスクも存在し，さらに Grad-CAM も勾配が不安定であることから，学習の妨げとなる可能性がある．

そこで本研究では，学習初期には視線一貫性損失の寄与度 λ を小さく（あるいは 0 に）設定し，学習が進むにつれてにその影響を強めるカリキュラム学習的手法を採用する．

具体的には，学習エポック t における視線一貫性損失の寄与度 λ_t を次式で定義する．

$$\lambda_t = \begin{cases} 0 & (t < T_0) \\ \lambda_{\max} \frac{t-T_0}{T_1-T_0} & (T_0 \le t < T_1) \\ \lambda_{\max} & (T_1 \le t) \end{cases} \tag{18}$$

ここで，T_0 は視線一貫性損失の導入開始エポック，T_1 は最大寄与度 $\lambda_{\max}$ に到達するエポックを表す．

このように，学習初期には視線一貫性損失の影響を抑制し，モデルが基本的な特徴抽出能力を獲得した後に徐々に制約を強化することで，分類性能の低下を防ぎつつ，視線一貫性の向上を図ることが可能となる．

4 実験設定

4.1 データセット

本研究では，視線一貫性指標の有効性を異なるドメインにおいて検証するため，以下の 2 種類の画像データセットを用いる．

- Brain Tumor MRI Dataset [11]
- Chest X-Ray Image (Pneumonia) [12]

Brain Tumor MRI Dataset（以下, Brain Tumor MRI）および Chest X-Ray Image (Pneumonia)（以下，Chest X-Ray）は，いずれも医用画像を対象とした分類タスクとして広く用いられている．

Brain Tumor MRI は glioma, meningioma, pituitary, no tumor の 4 クラスから構成される多クラス分類問題であり，Chest X-Ray は肺炎の有無を判別する 2 クラス分類問題である．

一貫性指標は，予測の際に画像の特定の領域を注視する必要があるタスクにおいて特に有効であると考えられるため，医用画像データセットを選定した．レントゲン画像や MRI 画像は，撮影方法が標準化されており，また患者の解剖学的構造が類似しているため，同一クラスに属する画像に対してモデルが一貫した注目領域を持つことが期待される．

4.2 前処理

ここでは，各データセットの画像特性および分類タスクの性質を考慮し，前処理およびデータ拡張を設計した．

Brain Tumor MRI および Chest X-Ray は医用画像を対象としており，色情報が診断上本質的でないことから，全ての画像をグレースケール（1 チャネル）に変換した．

学習時には，過学習の抑制および視線分布のロバスト性向上を目的として，Random Horizontal Flip および Random Resized Crop による軽度のデータ拡張を適用した．特に Chest X-Ray では元画像のサイズのばらつきが大きいため，リサイズ前にランダムクロップを行うことで，スケール差や画像周辺部の余白の影響を低減した．評価時には，データ拡張は行わず，リサイズおよび正規化のみを適用した．

表 1: モデル設計レベルの概要

Level 1	最小ベースライン （CAM の最低限挙動）
Level 2	標準的 CNN （現実的ベースライン）
Level 3	Residual 構造 （勾配安定性）
Level 4	Attention 機構 （注視領域の集中）
Level 5	高表現力モデル （精度と説明性のトレードオフ）

全ての画像は画素値を[0,1]にスケーリングした後，チャネル数に応じて平均 0.5，標準偏差 0.5 による正規化を行った．これにより，異なるデータセット間で入力分布のスケールを揃え，学習の安定性を確保した．

リサイズについては，Brain Tumor MRI および Chest X-Ray ともに，224 × 224 ピクセルにリサイズを行った．

なお，本研究における視線一貫性損失および評価指標は正規化空間上で定義されているため，入力解像度や前処理の差異によるスケールの影響を受けない．今回は同じ解像度を用いたが，本来は異なる解像度であっても問題はなく，各データセットの特性に応じた前処理を適用することが可能である．

4.3 モデル構成

本研究では，視線一貫性指標の性質を，多様なモデル条件下において体系的に検証するため，構造および表現能力の異なる 5 種類の畳み込みニューラルネットワークを用いる．これらのモデルは，Grad-CAM による視線抽出を前提として設計されており，いずれも最終畳み込み層をフック層として Grad-CAM を算出する．

各モデルは，ネットワークの深さ，表現容量，および内部構造の違いが視線一貫性および分類性能に与える影響を段階的に比較することを目的としている．モデルの設計レベルと主な比較観点を表 1 に示す．

ここでの Level とは，値が大きくなるほどモデルの構造的複雑性および表現能力が増加することを示す．ただし，これは必ずしも分類性能の単調な向上を意味するものではなく，各レベルにおける視線一貫性の挙動を比較することが主目的である．

以下では，各モデルの設計意図を簡潔に述べる．なお，各モデルの具体的な構造やパラメータ設定については，付録の図 appx.1〜図 appx.5 に詳細を示す．

4.3.1 Level 1:最小ベースライン（CAM の最低限挙動）

Level 1 は, 最小限の畳み込み層から構成される単純な CNN であり，Grad-CAM によって得られる視線分布の最下限的な挙動を確認するためのベースラインとして位置づける．

本モデルは計算量および表現能力の両面で最も小さく，分類精度は限定的である一方，Grad-CAM が示す注目領域が, 最小構成のモデルにおいてどの程度まで意味を持ち得るかを評価する基準となる．

4.3.2 Level 2:標準的 CNN（現実的ベースライン）

Level 2 は，Batch Normalization および複数段の畳み込み・プーリング層を備えた，一般的な CNN 構成である．

本モデルは, 実用的な分類性能と Grad-CAM の可視化品質のバランスを取ったベースラインとして用いる. 以降のモデルとの比較において, 精度および視線一貫性の基準点として機能する．

4.3.3 Level 3: Residual 構造（勾配安定性）

Level 3 では，Residual Block を導入することでネットワークをより深くし，勾配消失問題の緩和および学習安定性の向上を図る．

Residual 構造は分類性能の向上だけでなく, Grad-CAM において用いられる勾配情報の安定性にも影響を与える可能性がある. 本モデルでは, 深層化による視線分布の変化や一貫性への影響を検証する．

4.3.4 Level 4: Attention 機構（注視領域の集中）

Level 4 では, Squeeze-and-Excitation (SE) ブロックを用いたチャネル方向の Attention 機構を導入する．

Attention 機構により, モデルは「どの特徴チャネルに注目するか」を明示的に学習することが可能となる. 本モデルでは, 注視領域がより局所的かつ集中した場合に，視線一貫性指標がどのように変化するかを評価する．

4.3.5 Level 5:高表現力モデル（精度と説明性のトレードオフ）

Level 5 は，Residual 構造を基盤としてネットワークの深さおよび幅を拡張した，高表現力モデルである．

本モデルは高い分類性能が期待される一方, Grad-CAM による注目領域が分散しやすくなる可能性がある. ここでは, 分類精度の向上と視線一貫性（説明性）との間に存在するトレードオフを検証する目的で本モデルを導入する．

4.3.6 一貫性指標のモデル依存性について

なお, 各モデルにおいて Grad-CAM を算出するフック層の空間解像度（すなわち CAM のサイズ）は，ネットワーク構造や深さに応じて異なる．

しかし，本研究で提案する視線一貫性指標は, Grad-CAM を空間的に正規化した後，クラス内平均との差分布として定義されているため，フック層の空間解像度や CAM のサイズに依存しない. そのため, 異なるモデル構造間においても, 視線一貫性を同一の尺度で定量的に比較することが可能である．

4.4 学習条件

本研究では，各データセットおよびモデル構成に対して, 共通の学習設定を基本としつつ, 安定した学習が行えるよう最小限の調整を行った．前述の通り，本研究では,

1. 学習済みモデルに対する解析的評価としての一貫性評価
2. モデル学習時に一貫性を誘導するための損失関数としての一貫性制約

の両面から視線一貫性指標の有効性を検証するため，各実験における学習条件を以下に示す．

4.4.1 最適化手法

最適化手法には Adam オプティマイザを用いた. Adam は学習率の自動調整機構を備えており, 異なるモデル規模やデータセット間においても比較的安定した収束が得られることから，本研究の目的に適している．

- 最適化手法: Adam
- $\beta_1 = 0.9, \beta_2 = 0.999$
- $\varepsilon = 1 \times 10^{-8}$

4.4.2 学習率およびスケジューリング

初期学習率は全ての実験において共通とし，以下の値を用いた．

- 初期学習率: $\eta_0 = 1 \times 10^{-4}$

学習の進行に伴う過学習の抑制および収束安定性の向上を目的として，学習率スケジューラを導入した．具体的には，検証損失が一定エポック改善しない場合に学習率を減衰させる ReduceLROnPlateau を用いた．

- 学習率減衰率: 0.1
- patience: 5 epoch

4.4.3 損失関数

分類タスクはいずれも多クラス分類として定式化し，損失関数には Cross Entropy Loss を用いた．2 クラス分類の場合も，ラベルを one-hot ベクトルに変換し, 多クラス分類の枠組みで扱う. これは，視線一貫性指標および損失関数の定義を一貫させるためである. なお, 2 クラス分類を多クラス分類として扱うことは数学的に等価であり，モデルの表現能力や学習挙動に影響を与えない．

学習済みモデルに対する解析的評価としての一貫性評価を行う場合，損失関数は単に Cross Entropy Loss L_{cls} のみを用いる．

一方，モデル学習時に一貫性を誘導するための損失関数としての一貫性制約を導入する場合，

3 章 5 節 1 項 で定義した学習用視線一貫性損失 $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}}$ を正則化項として加算する．

最終的な損失関数は以下のように定義される．

$$\mathcal{L} := L_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}} \tag{19}$$

ここで λ は視線一貫性損失の寄与度を制御するハイパーパラメータであり，本研究では以下の 5 種類の値を用いて実験を行った．

$$\lambda = \begin{cases} -100 & (\text{視線多様性促進}) \\ 0 & (\text{通常の分類学習}) \\ 50 & (\text{視線一貫性促進: 中程度}) \\ 100 & (\text{視線一貫性促進: 強度}) \\ 500 & (\text{視線一貫性促進: 非常に強度}) \end{cases} \tag{20}$$

ただし，今回はカリキュラム学習を採用しているため，実際の学習における視線一貫性損失の寄与度はエポック数に応じて変化する．したがって，上記の λ は最大寄与度 $\lambda_{\max}$ を表す．カリキュラム学習の詳細については 4 章 4 節 6 項 で述べる．

4.4.4 学習エポック数およびバッチサイズ

各実験における学習エポック数およびバッチサイズは以下の通りとした．

1. 学習済みモデルに対する解析的評価としての一貫性評価
   - 学習エポック数: 20
   - バッチサイズ: 32
2. モデル学習時に一貫性を誘導するための損失関数としての一貫性制約
   - 学習エポック数: 50
   - バッチサイズ: 32

通常の学習では，エポック数を 20 としたが，これは多くのモデルがこのエポック数で十分に収束するためである．一方，一貫性による正則化を導入する場合，学習が安定するまでに通常より多くのエポック数を要することから，学習エポック数を増加させた．

なお，全てのモデルおよびデータセットにおいて，同一の学習回数およびバッチサイズを用いることで，モデル構造の違いが視線一貫性および分類性能に与える影響を公平に比較できるよう配慮した．

4.4.5 視線一貫性損失の安定化および EMA クラス CAM の設定

視線一貫性損失は，同一クラスに属し，かつ正しく分類されたサンプル集合に基づいて算出されるため，バッチ内のサンプル数が少ない場合に不安定になりやすい．この問題を緩和するため，本研究ではクラスごとの CAM を指数移動平均（Exponential Moving Average; EMA）により平滑化した EMA クラス CAM を導入した．

EMA クラス CAM の更新は，対応するクラスに属する正解サンプルがバッチ内に存在する場合にのみ行い，初期値は最初に観測された当該クラスの CAM とした．EMA の減衰率は全実験で共通とし，以下の値を用いた．

- EMA 減衰率: $\alpha = 0.9$

4.4.6 カリキュラム学習の設定

視線一貫性損失は，学習初期に強い制約を課すと分類性能が低下する可能性があり，さらに Grad-CAM 自体も勾配が不安定であることから，学習の妨げとなるリスクが存在する．この問題を回避するため，視線一貫性損失の学習への導入は，カリキュラム学習的手法を用いて段階的に行った．

具体的には，学習エポック数 T_0 までは視線一貫性損失の寄与度を 0 とし，T_0 以降 T_1 までの間に徐々に最大寄与度 $\lambda_{\max}$ に到達するように設定した．

- 視線一貫性損失導入開始エポック: $T_0 = 2$
- 最大寄与度到達エポック: $T_1 = 10$

4.4.7 正則化および初期化

Level 2 以降のモデルでは，過学習の抑制を目的として，全結合層に対して Dropout を適用した．

- Level 2 ~ 4 Dropout 率: 0.5
- Level 5 Dropout 率: 0.6

また，畳み込み層および全結合層の重みは，PyTorch のデフォルト初期化（Kaiming 初期化）を用いた．

4.4.8 学習および評価の手順

学習は訓練データに対してのみ行い，各エポック終了時に検証データを用いて分類精度 Acc ，κ および視線一貫性指標 $\mathcal{L}_{\mathrm{CAM}}$ を算出した．

5 実験結果

5.1 視線一貫性と分類性能の関係（解析的評価）

本節では，学習済み CNN モデルに対して，分類性能（Cohen's κ 係数）κ と視線一貫性指標 $\mathcal{L}_{\mathrm{CAM}}$ の関係を解析的に評価する．

5.1.1 実験設定の要約

- 各モデルについて，学習過程の各エポック終了時に同一の検証データセットを用いて評価を行った．
- 各エポックにおいて，以下を算出した．
  - 分類性能：Cohen's κ 係数 κ
  - 視線一貫性：$\mathcal{L}_{\mathrm{CAM}}$
- κ を横軸， $\mathcal{L}_{\mathrm{CAM}}$ を縦軸とした散布図を作成し，モデルごとに色分けして可視化した．
- 同一条件下で異なる乱数シードを用いて 6 回の独立試行を行い，各モデルについて κ と $\mathcal{L}_{\mathrm{CAM}}$ の 相関係数（Pearson） を算出した．

評価は以下の 2 データセットで行った．

- Brain Tumor MRI
- Chest X-Ray

表 2: κ 係数 と 一貫性損失$\mathcal{L}_{\mathrm{CAM}}$ の
平均相関係数（$n = 6$）

Level	Correlation($\kappa, \mathcal{L}_{\mathrm{CAM}}$)	
	Brain Tumor MRI	Chest X-Ray
1	0.735	−0.5587
2	−0.6263	−0.5901
3	−0.4495	−0.0956
4	−0.6669	0.0062
5	−0.6309	−0.325

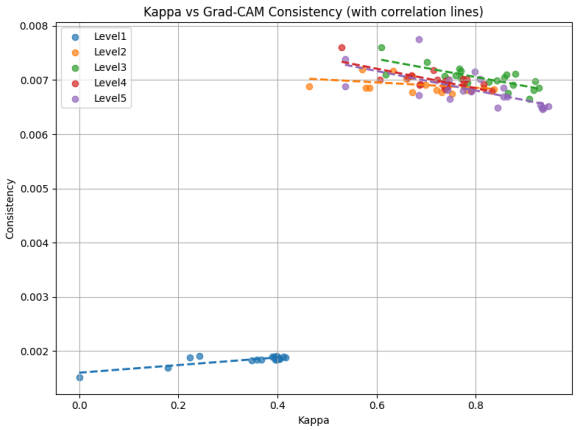


図 1: Brain Tumor MRI における
$\kappa - \mathcal{L}_{\mathrm{CAM}}$ 散布図の例

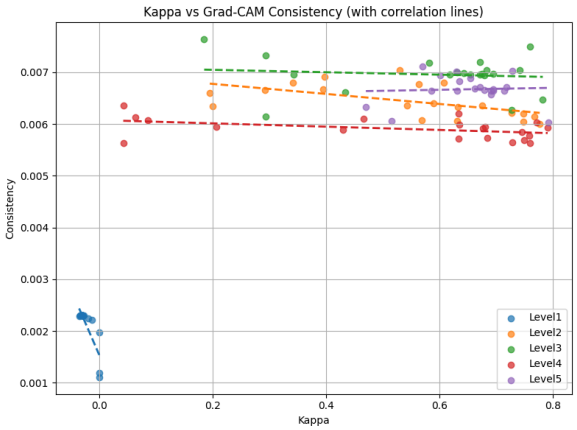


図 2: Chest X-Ray における
$\kappa - \mathcal{L}_{\mathrm{CAM}}$ 散布図の例

5.1.2 相関係数の定量的結果

表 2 に，6 回の独立試行における κ 係数 と 一貫性損失$\mathcal{L}_{\mathrm{CAM}}$ の平均相関係数を示す．

5.1.3 散布図による傾向の可視化

図 1（Brain Tumor MRI），図 2（Chest X-Ray）に，6 回の試行のうち代表的な 1 回における $\kappa - \mathcal{L}_{\mathrm{CAM}}$ 散布図を示す．各図には最小二乗法による近似直線を併記している．

これらの図から，以下の傾向が確認できる．

- 多くのモデルでは，弱〜中程度の負の相関あるいはほぼ無相関が観測された
- Brain Tumor MRI においては，ほとんどのモデルで負の相関が顕著であった（$-0.4 \sim -0.6$ 程度）
- Chest X-Ray においては，比較的弱い負の相関が観測されたモデルが多く，一部のモデルではほぼ無相関であった（$-0.006 \sim 0.6$ 程度）
- 弱〜中程度の負の相関あるいはほぼ無相関が観測された一方で，一部のモデルでは，正の相関を示すモデルも存在した
- Level 1 は，κ と $\mathcal{L}_{\mathrm{CAM}}$ の両方において，散布図上で，他のモデルとは異なる領域に分布する傾向があった

5.1.4 視線一貫性指標のスケール

理論的には，正規化済み Grad-CAM に基づく $\mathcal{L}_{\mathrm{CAM}}$ は $[0, 1]$ の範囲を取ると解釈可能であるが，実際の実験では，その値は 10^{-3} オーダーに集中していた．

例えば，

- $\kappa \approx 0.9$ の高性能モデルにおいても $\mathcal{L}_{\mathrm{CAM}} \approx 6 \times 10^{-3}$
- $\kappa \approx 0.2$ の低性能モデルにおいても $\mathcal{L}_{\mathrm{CAM}} \approx 1.5 \times 10^{-3}$

といった結果が観測された．

5.2 視線一貫性正則化の学習効果（学習的評価）

本節では，視線一貫性損失を損失関数に組み込んで学習を行った場合の，分類性能（正答率 Acc）の変化を評価する．

5.2.1 実験設定の要約

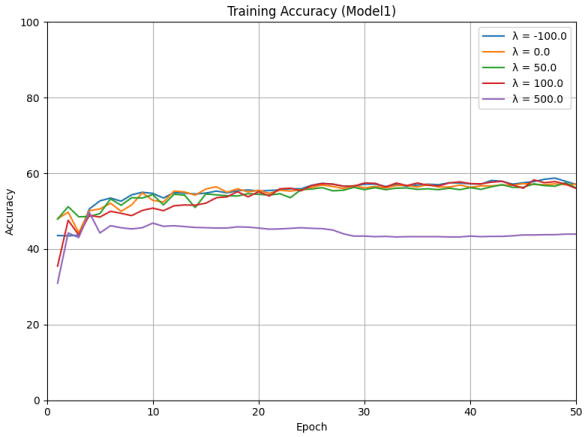
- 各モデルについて，5 種類の視線一貫性損失寄与度 λ を用いて学習を行った．
  - $\lambda \in \{-100, 0, 50, 100, 500\}$
- 各学習条件において，同一の検証データセットを用いて評価を行った．
- 各エポックにおいて，以下を算出した．
  - 分類性能：正答率 Acc
- エポック ごとの Acc をプロットし，各モデル・各 λ ごとに学習曲線を可視化した．

評価は以下の 2 データセットで行った．

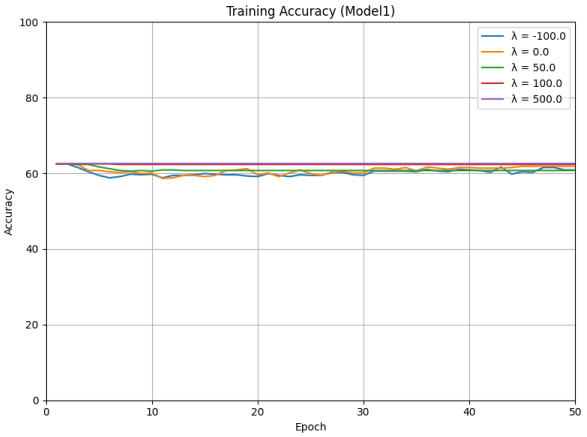
- Brain Tumor MRI
- Chest X-Ray

5.2.2 学習曲線の可視化

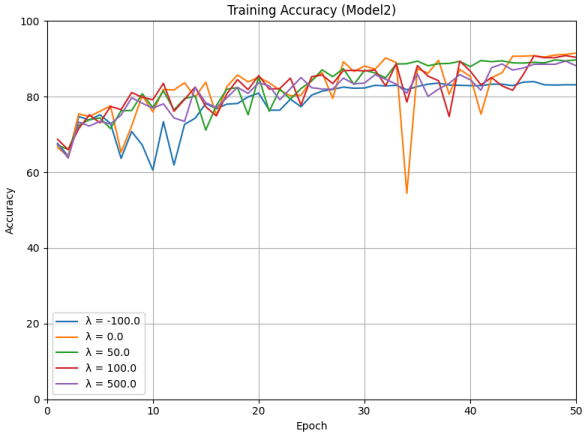
学習曲線を 図 4 に示す．グラフはデータセット・モデルごとに作成し，各 λ ごとに色分けしている．各グラフの横軸はエポック数，縦軸は正答率 Acc を表す．



(a) Brain Tumor MRI - Level 1 モデル

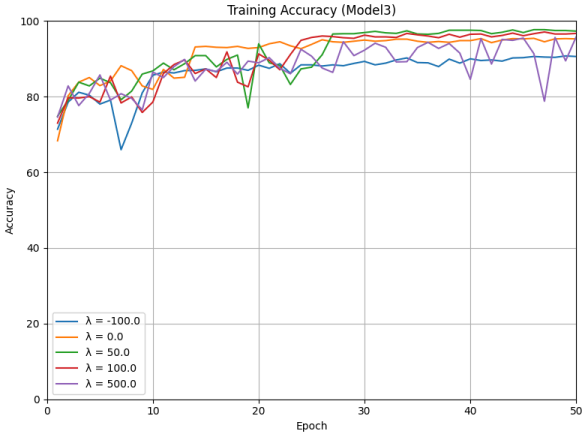


(b) Chest X-Ray - Level 1 モデル

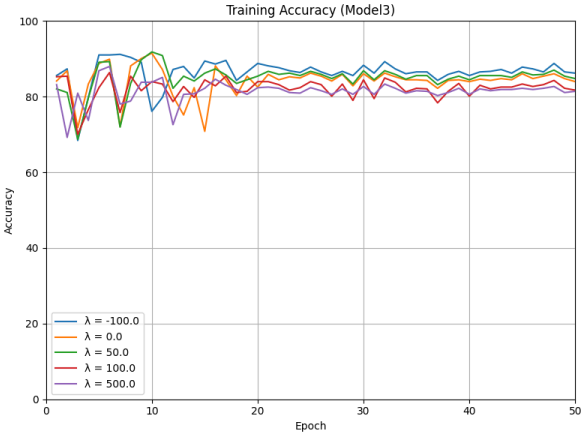


(c) Brain Tumor MRI - Level 2 モデル

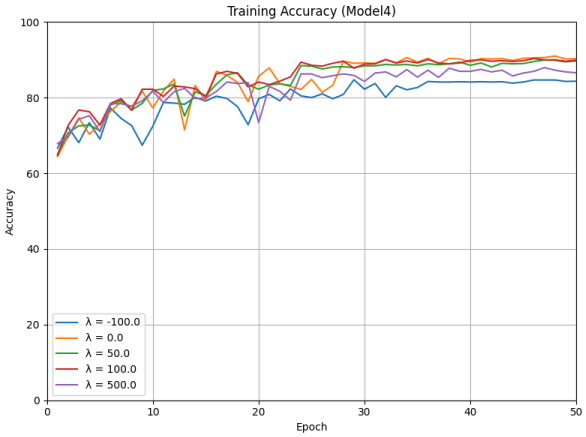
(d) Chest X-Ray - Level 2 モデル



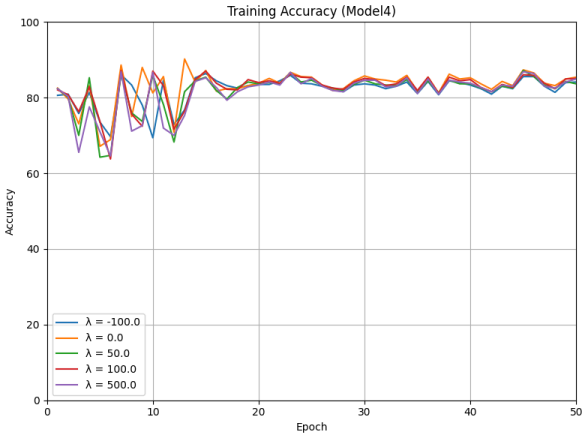
(e) Brain Tumor MRI - Level 3 モデル



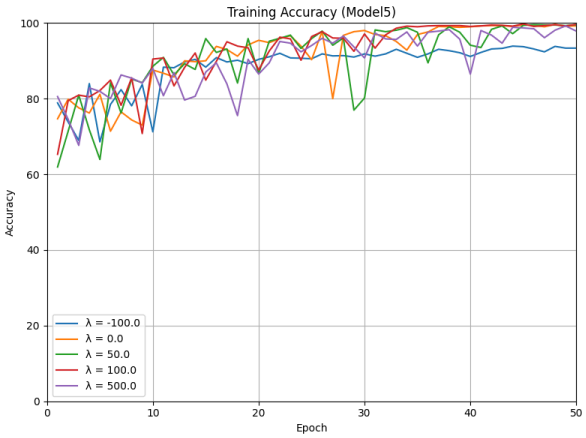
(f) Chest X-Ray - Level 3 モデル



(g) Brain Tumor MRI - Level 4 モデル



(h) Chest X-Ray - Level 4 モデル



(i) Brain Tumor MRI - Level 5 モデル

(j) Chest X-Ray - Level 5 モデル

図 4: 各データセット・各モデルにおける正答率 Acc の学習曲線
（横軸：エポック数，縦軸：正答率 Acc | λ ごとに色分け）

これらの図から，以下の傾向が確認できる．

- λ 値による分類性能 Acc の違いの傾向は，データセットで大きく異なった．
- Brain Tumor MRI では以下のような傾向が見られた
  - 多くのモデルでは，λ が 50 または 100 のとき，分類性能 Acc が向上する傾向が観測された
  - 一方，$\lambda = 500$ のように極端に大きな値では，学習が不安定になり，分類性能が低下する場合があった．この傾向は特に表現力の低いモデル（Level 1）で顕著であった
  - $\lambda = -100$ とした際に，分類性能 Acc は最も低下する傾向が見られた
  - 総合的には，$-100 \to 0, 500 \to 50, 100$ の順に分類性能 Acc が高い傾向が見られた
- Chest X-Ray では以下のような傾向が見られた
  - 正答率は，エポックごとに多少の変動が見られるが[5]，この変動パターンが λ ごとに大きく異なる傾向は見られなかった
  - Brain Tumor MRI と比べると，各 λ による分類性能 Acc の差異は小さい傾向が見られた

6 考察

6.1 視線一貫性と分類性能の関係（解析的評価の解釈）

本研究では，「分類性能が高いモデルほど，Grad-CAM に基づく視線が一貫する」という仮説を立て，Cohen's κ 係数と視線一貫性損失 $\mathcal{L}_{\mathrm{CAM}}$ の関係を解析的に評価した．

実験の結果，明確な正の相関はほとんど観測されず，多くの条件において弱〜中程度の負の相関，あるいはほぼ無相関が確認された．特に Brain Tumor MRI では，複数のモデルで一貫して負の相関が観測され，分類性能の高いモデルほどクラスごとに安定した注目領域を形成している可能性が示唆された．一方で，Chest X-Ray では，このような傾向は弱く，モデル間で一貫した関係は確認されなかった．

以上より，本仮説はすべての条件において強く支持されるとは言えないものの，データセットによっては部分的に支持されると結論づけられる．

6.1.1 視線一貫性指標のスケールに関する考察

一貫性損失 $\mathcal{L}_{\mathrm{CAM}}$ は理論的には $[0, 1]$ の範囲を取り得るが，実際には 10^{-3} オーダーの非常に小さい値に集中した．この要因として，以下が考えられる．

- Grad-CAM において理論的最大差分に相当する状況が現実的でないこと
- 画素単位の二乗差分と空間平均による値の縮小効果
- 同一モデル・同一データセット内で CAM が構造的に類似する傾向を持つこと

$\mathcal{L}_{\mathrm{CAM}} = 1$ になる場合とは，2 つの CAM 画像が，全画素において，一方が 0，他方が 1 となるような極端な場合である．このような状況は，現実的にはほとんど発生しないため，実験においても $\mathcal{L}_{\mathrm{CAM}}$ は非常に小さい値に留まったと考えられる．

この結果，定義としては妥当である一方，人間の直感や可視化との対応づけは容易ではないスケールとなっている．この点は，非線形変換や別種の距離指標を用いることで改善の余地がある．

6.1.2 CAM 解像度の影響と Level 1 の特異性

解析的評価において，Level 1 は分類性能・視線一貫性の双方が低いにもかかわらず，他モデルとは異なる分布を示した．調査の結果，Level 1 の CAM 解像度が他モデルよりも高いことが確認された．

このことは，CAM の空間解像度が異なる場合，正規化後であっても視線一貫性指標の統計的性質が変化する可能性を示している．本研究では CAM 解像度を統一していないため，解像度非依存な評価指標の設計は今後の課題である．

6.2 視線一貫性正則化の学習効果（学習的評価の解釈）

視線一貫性損失を正則化項として学習に組み込んだ実験では，データセット間で大きく異なる挙動が観測された．

Brain Tumor MRI では，適度な寄与度（$\lambda = 50, 100$）において分類性能が向上する傾向が複数のモデルで確認された．この結果は，解析的評価で観測された「高性能モデルほど視線が安定する」という傾向と整合的であり，視線一貫性正則化が学習を補助する可能性を示している．

一方で，$\lambda = 500$ のように過度に強い正則化では学習が不安定になり，特に表現力の低いモデルにおいて性能低下が顕著であった．このことから，視線一貫性は有用である一方，分類損失とのバランスが重要であることが示唆される．

対照的に，Chest X-Ray では，λ の違いによる分類性能の差異は小さく，視線一貫性正則化の明確な効果は確認されなかった．

6.3 データセット依存性に関する考察

本研究の結果は，Grad-CAM に基づく視線一貫性がデータセット依存的な性質を持つことを示唆している．

Brain Tumor MRI 画像では，クラスごとに腫瘍が局所的かつ比較的一貫した空間構造を持つため，「どの領域を注目すべきか」がクラス内で共有されやすい．このような条件では，視線一貫性という制約がモデルの学習に有益に働く可能性がある．

[5] Chest X-Ray はテスト用の画像数が 600 程度と少なく，評価時のバッチ内サンプル数の影響を受けやすいため，エポックごとに多少の正答率の変動が見られる．この特徴自体は Brain Tumor MRI でも同様である．

一方，Chest X-Ray 画像では，病変が肺野全体に広がる場合や，位置・形状のばらつきが大きい場合が多く，クラス内で注目領域を一意に定義することが難しい．このようなタスクでは，「同一領域への注目」を仮定する視線一貫性正則化は必ずしも適合せず，学習効果が限定的であったと考えられる．

重要なのは，これらの違いを医学的診断基準としてではなく，画像構造とタスク特性の違いとして捉える点である．

6.4 本研究の示唆と限界

本研究は，Grad-CAM に基づく視線一貫性を定量化し，解析的・学習的の両側面から評価した点に意義がある．一方で，CAM 解像度の違いや指標スケールの問題，学習初期における不安定性など，いくつかの制約も明らかとなった．

今後は，CAM 解像度に依存しない指標設計や，学習後半のみで視線一貫性を導入するカリキュラム学習の検討が重要な課題である．

7 結論

本研究では，Grad-CAM に基づく視線一貫性指標を定義し，CNN における分類性能との関係を解析的および学習的に評価した．

解析的評価では，データセット Brain Tumor MRI Dataset において，分類性能と視線一貫性指標の間に中程度の負の相関が観測され，高性能モデルほどクラスごとに安定した注目領域を形成する傾向が確認された．一方，データセット Chest X-Ray（Pneumonia）では，このような明確な関係は観測されなかった．

さらに，学習的評価では，Brain Tumor MRI Dataset において，視線一貫性損失を適度な強さで正則化項として導入することで，分類性能が向上する傾向が確認された．これに対し，Chest X-Ray Image（Pneumonia）では，視線一貫性正則化による性能向上は明確には確認されなかった．

これらの結果から，Grad-CAM に基づく視線一貫性正則化は，データセットに依存して有効性が大きく異なる手法であり，「同一クラスにおいて注目領域が比較的一貫して定義可能なタスク」において有効である可能性が示唆された．

本研究は，説明可能性指標を単なる可視化に留めず，学習の補助情報として活用する可能性を示した点に意義がある．

謝辞

本研究を進めるにあたり，終始にわたって御指導・御助言を賜りました，有明工業高等専門学校 創造工学科 人間・福祉工学系 Gauthier Lovic 教授に深く感謝いたします．

参考文献

[1] A. Krizhevsky, I. Sutskever, と G. E. Hinton, 「ImageNet Classification with Deep Convolutional Neural Networks」, *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, と K. Weinberger, 編, Curran Associates, Inc., 2012, p. . [Online]. 入手先: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[2] K. Simonyan と A. Zisserman, 「Very Deep Convolutional Networks for Large-Scale Image Recognition」. [Online]. 入手先: https://arxiv.org/abs/1409.1556

[3] K. He, X. Zhang, S. Ren, と J. Sun, 「Deep Residual Learning for Image Recognition」. [Online]. 入手先: https://arxiv.org/abs/1512.03385

[4] M. Bellucci, N. Delestre, N. Malandain, と C. Zanni-Merk, 「Towards a terminology for a fully contextualized XAI」, *Procedia Computer Science*, vol. 192, pp. 241–250, 2021, doi: https://doi.org/10.1016/j.procs.2021.08.025.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, と A. Torralba, 「Learning Deep Features for Discriminative Localization」. [Online]. 入手先: https://arxiv.org/abs/1512.04150

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, と D. Batra, 「Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization」, *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 10 月 2019, doi: 10.1007/s11263-019-01228-7.

[7] A. Binder, G. Montavon, S. Bach, K.-R. Müller, と W. Samek, 「Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers」. [Online]. 入手先: https://arxiv.org/abs/1604.00825

[8] M. Sundararajan, A. Taly, と Q. Yan, 「Axiomatic Attribution for Deep Networks」. [Online]. 入手先: https://arxiv.org/abs/1703.01365

[9] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, と B. Kim, 「Sanity Checks for Saliency Maps」. [Online]. 入手先: https://arxiv.org/abs/1810.03292

[10] S. Hooker, D. Erhan, P.-J. Kindermans, と B. Kim, 「A Benchmark for Interpretability Methods in Deep Neural Networks」. [Online]. 入手先: https://arxiv.org/abs/1806.10758

[11] M. Nickparvar, 「Brain Tumor MRI Dataset」. [Online]. 入手先: https://www.kaggle.com/dsv/2645886

[12] D. Kermany, K. Zhang, と M. Goldbaum, 「Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification」. [Online]. 入手先: https://doi.org/10.17632/rscbjbr9sj.2

[13] A. Paszke ほか, 「PyTorch: An Imperative Style, High-Performance Deep Learning Library」. [Online]. 入手先: https://arxiv.org/abs/1912.01703

[14] E. Remy, 「neural-netz, a Typst Package」. [Online]. 入手先: https://hal.science/hal-05401124

付録

A 使用ライブラリおよび実行環境

本研究で用いた実験環境および主要なソフトウェア依存関係を以下に示す．これらの情報は，本研究の実験結果の再現性を担保する目的で記載する．

A.1 実行環境

実験は，以下の計算機環境上で実施した．

- OS: Ubuntu 24.04.3 LTS (x86_64)
- Kernel: 6.8.0-90-generic
- CPU: Intel Core i7-9700K (8 cores, up to 4.90 GHz)
- GPU: NVIDIA GeForce RTX 2080 Mobile

GPU を用いた計算は CUDA 対応環境上で行い，深層学習モデルの学習および Grad-CAM の算出を高速化している．

A.2 Python 実行環境

- Python バージョン: Python 3.11.14
- Python 要件: `Python >= 3.11`

A.3 使用ライブラリ

本研究で使用した主な Python ライブラリおよびそのバージョンは以下の通りである．

- PyTorch: `torch >= 2.9.0` [13]
- TorchVision: `torchvision >= 0.24.0`
- NumPy: `numpy >= 2.3.4`
- Pandas: `pandas >= 2.3.3`
- Matplotlib: `matplotlib >= 3.10.7`
- tqdm: `tqdm >= 4.67.1`

データセットの取得や実験管理のため，以下の補助的ライブラリを使用した．

- kaggle: `kaggle >= 1.8.3`
- kagglehub: `kagglehub >= 0.3.13`
- gdown: `gdown >= 5.2.0`
- requests: `requests >= 2.32.5`
- python-dotenv: `python-dotenv >= 1.2.1`
- ipython: `ipython >= 9.9.0`

開発時のコード品質管理には，以下のツールを使用した．

- ruff: `ruff >= 0.14.10`

B 実装の公開リポジトリ

本研究で使用した実装は，再現性および検証可能性の確保を目的として，GitHub 上で公開している．

- リポジトリ URL: https://github.com/59GauthierLab/Yonemura_Research_Public

本リポジトリには，以下の内容が含まれている．

- 各モデル構成（Level 1〜5）の実装
- Grad-CAM の算出および正規化処理
- 視線一貫性損失および評価指標の実装
- 各データセットに対応した前処理および学習スクリプト
- 実験結果およびその結果を再現するためのパラメータ記録

なお，リポジトリ内のコードは，付録 A に示した実行環境およびライブラリ構成を前提として動作確認を行っている．

C 各モデルのネットワーク構成図

本研究で用いた各モデルのネットワーク構成図を図 appx.1 〜図 appx.5 に示す（次ページ参照）．

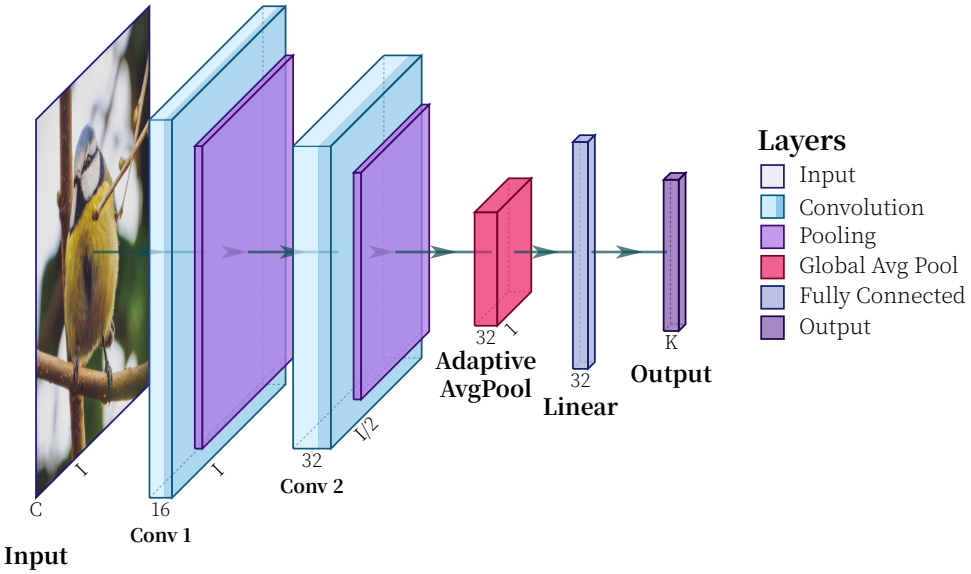


図 appx.1: Level 1（最小ベースライン）モデルのネットワーク構成 [14]

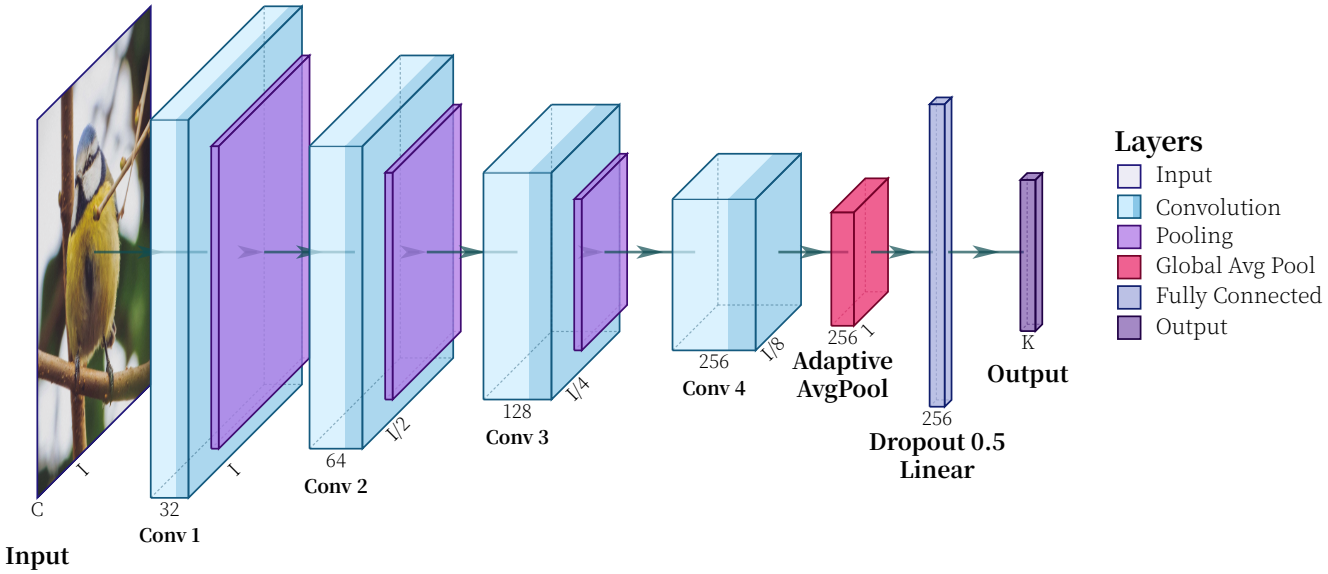


図 appx.2: Level 2（標準的 CNN）モデルのネットワーク構成 [14]

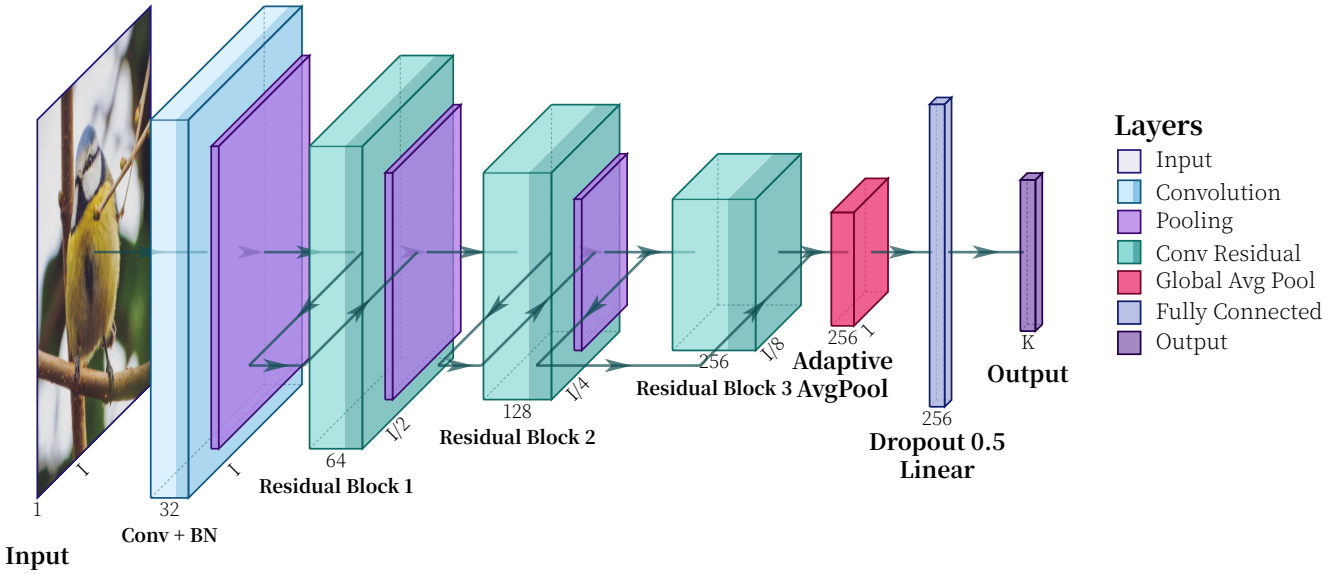


図 appx.3: Level 3（Residual 構造）モデルのネットワーク構成 [14]

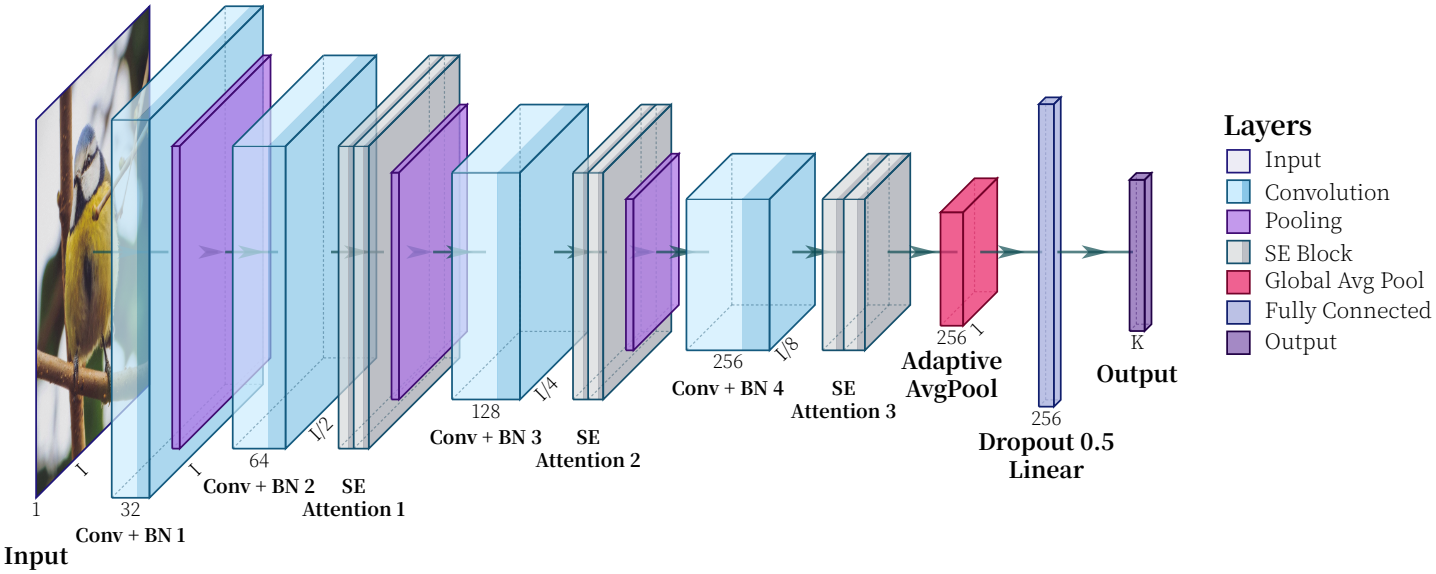


図 appx.4: Level 4（Attention 機構）モデルのネットワーク構成 [14]

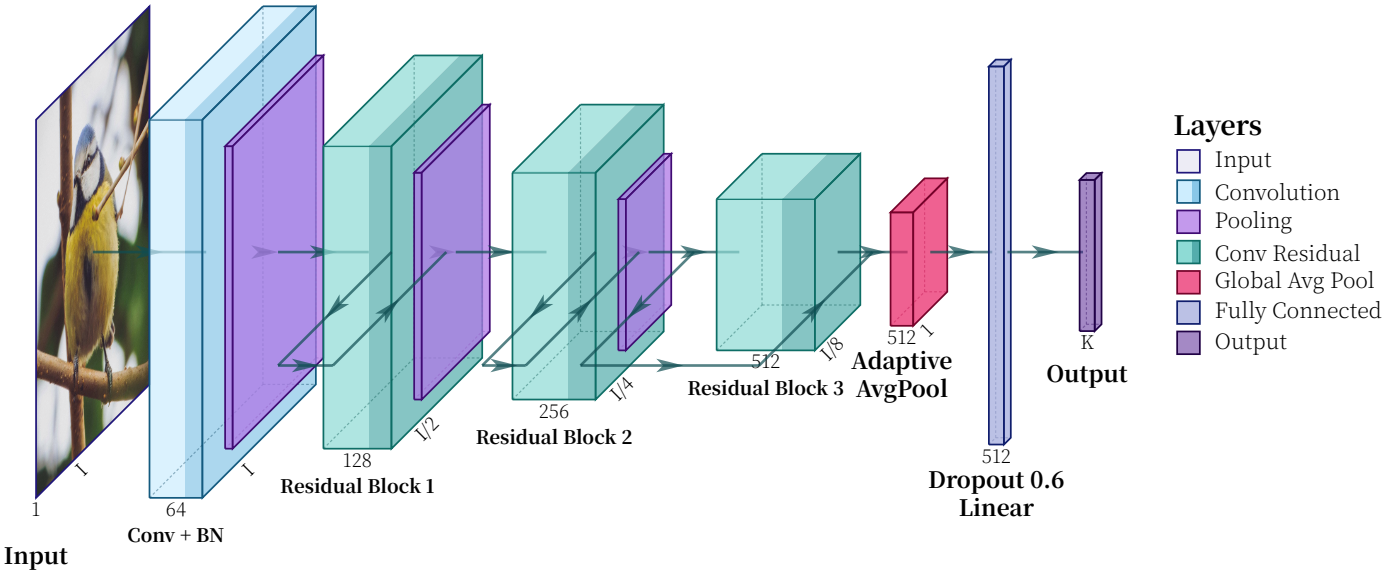


図 appx.5: Level 5（高表現力）モデルのネットワーク構成 [14]