# 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

2026-02-18

# 1. 背景と目的

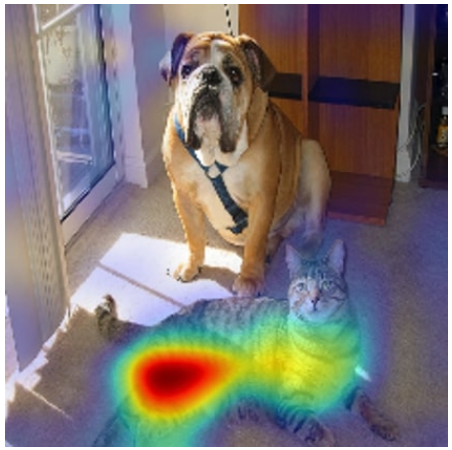
# 1.1 背景

- **Grad-CAM**: CNN の判断根拠を可視化するための手法
- どんな風に可視化する？
  - 予測に寄与した領域をヒートマップで提示

    → モデルが「**どこを見て判断したのか（視線）**」っぽく解釈できる

**例:「犬」「猫」を認識する CNN モデルで Grad-CAM 適用**

「犬」の判断根拠となる部分



「猫」の判断根拠となる部分

# 1.2 Grad-CAM の使われ方

- 画像系 AI の判断根拠の可視化に使われている Grad-CAM

> 「このモデルは（可視化結果を見た感じ）ちゃんと正しい部分を見て推論してそうだな → 良いモデルと判断

- しかし，
  - モデルの良さが **定性的** に議論されがち
  - モデルの良さの根拠が**後付け的**になりがち

  → 「評価・制御できる量」として扱う枠組みを作りたい

Grad-CAM の可視化結果を「モデルの視線」とみなし， **同一クラス内でどれだけ視線が揃うか** を **視線一貫性** として定義し，その妥当性を評価．

# 2. 提案：視線一貫性指標

# 2.1 指標の概要

**視線一貫性**: 同一クラスに属する複数入力に対して，モデルが空間的に類似した Grad-CAM 分布を示す性質 と定義

1. データセットから**モデルが正解したサンプル**を抽出
2. 正解したサンプルから，クラスごとに**平均 CAM マップ**を計算
3. 正解したサンプルの CAM と平均マップの**偏差(L2 距離)**を計算
4. 各サンプルの偏差を合計・正規化

各サンプルの CAM マップの分散的な量をつかって「**視線がばらついているのか，集中（一貫）しているのか**」測っているイメージ

# 3. 実験設計

# 3.1 実験の全体像

視線一貫性指標の評価は 2 本立て

1. **解析的評価**：学習済みモデルの各エポックで
   **分類性能** $\kappa$ と **視線一貫性** $\mathcal{L}_{\mathrm{CAM}}$ の関係を観察
2. **学習的評価**：**視線一貫性損失** $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}}$ による正則化項の重み $\lambda$
   を変えて分類性能がどう変わるかを観察

これら 2 つの実験を

2 つのデータセット　5 つのモデル　で検証

# 3.2 データセット（2 つ）

- **Brain Tumor MRI Dataset**
  - 脳の MRI 画像 → 脳腫瘍の病名
  - 4 クラス（glioma, meningioma, pituitary, no tumor）
- **Chest X-Ray Image (Pneumonia)**
  - 肺のレントゲン画像 → 肺炎の有無
  - 2 クラス（NORMAL, PNEUMONIA）

共通：

- グレースケール（1ch），224×224　・バッチサイズ: 32
- 学習：Random Horizontal Flip / Random Resized Crop（軽度）
- 評価：リサイズ＋正規化のみ

# 3.3 モデル（5 つ）

表現力の異なる CNN を **5 段階** で用意
→ 視線挙動がモデル構造でどう変わるかを見たいため

| モデル名 | 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（SE ブロック：注視の集中） |
| Level 5 | 高表現力（精度と説明性のトレードオフ） |

- Level が高いほど，表現力が高い
- 「表現力が高い≠ 分類性能が高い 」の点には注意

# 3.4 学習条件



| 項目 | 設定 |
|---|---|
| エポック数 | 解析的評価: 20, 学習的評価: 50 |
| 最適化手法 | Adam（$\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1 \times 10^{-8}$） |
| 学習率 | 初期学習率: $\eta_0 = 1 \times 10^{-4}$<br>スケジューラ: ReduceLROnPlateau ($r$ = 0.9, patience=5epoch) |
| 損失関数 | 解析的評価: $L = L_{\text{cls}}$ （$L_{\text{cls}}$: Cross Entropy Loss）<br>学習的評価: $\mathcal{L} = L_{\text{cls}} + \lambda \mathcal{L}_{\text{CAM}}^{\text{train}}$ |
| 一貫性重み | $\lambda_0 \in \{-100, 0, 50, 100, 500\}$, カリキュラム学習あり |
| 評価指標 | Cohen's κ 係数 $\kappa$， 正答率 $\mathrm{Acc}$ |
| 安定化手法 | EMA（$\alpha = 0.9$）, カリキュラム学習($\lambda$ を変化) |

# 4. 結果と考察

# 4.1 結果 1：解析的評価

| Lv. | 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 |

$\kappa - \mathcal{L}_{\mathrm{CAM}}$ の平均相関係数（6 回試行）
一貫性 $\uparrow \Leftrightarrow \mathcal{L}_{\mathrm{CAM}} \downarrow$ の点に注意

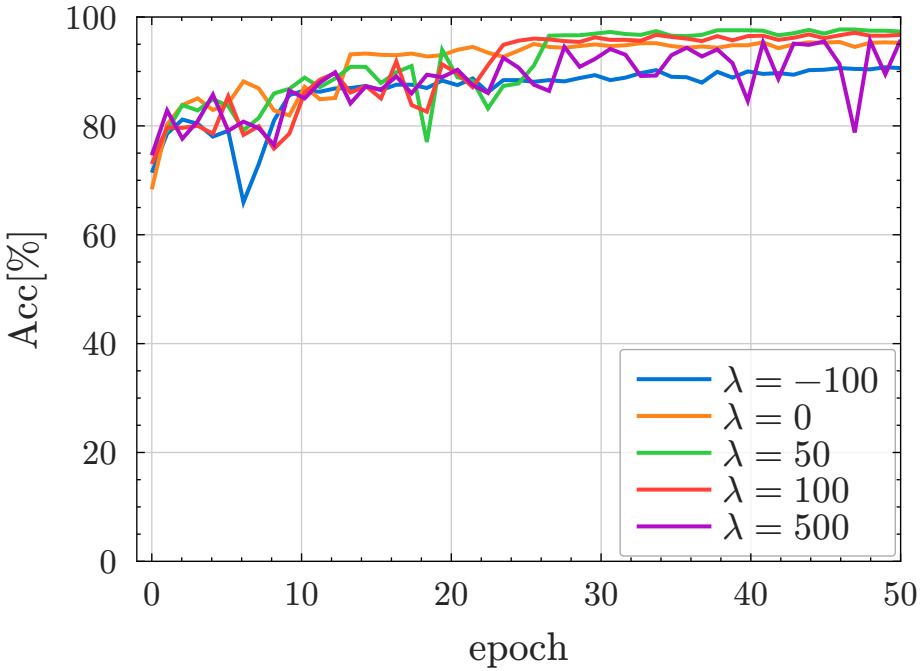
- 分類性能 $\kappa$ が高いほど，一貫性が高くなる( $\mathcal{L}_{\mathrm{CAM}}$ が小さくなる)
  $\rightarrow \kappa - \mathcal{L}_{\mathrm{CAM}}$ に負の相関が出れば一貫性指標を支持する結果に
- Brain Tumor MRI は中程度の負の相関
- Chest X-Ray は中〜弱程度の負の相関あるいはほぼ無相関
- 一部のモデルでは正の相関も

- Brain Tumor MRI では 負の相関 → 一貫性の有効性を支持できそう
- Chest X-Ray では傾向が弱い → 弱く指示できるが，微妙

# 4.2 結果 2：学習的評価

- $\lambda$ 系列別で学習曲線を作成
  $\rightarrow \lambda > 0$ 系列で，分類性能 Acc が高いと一貫性指標を支持する結果に

- Brain Tumor MRI では **中程度**（$\lambda = 50, 100$）で性能が上がるモデルが複数
- Chest X-Ray では $\lambda$ による差が小さく明確な改善は限定的




学習曲線：Brain Tumor MRI - Level3

Brain Tumor MRI では視線一貫性による正則化により性能が改善

# 4.3 考察:指標の有用性

主要な結論：**視線一貫性の有効性はデータセット依存**
→ **画像構造とタスク特性** を見極めて使うなら有用

- Brain Tumor MRI：
  - クラスごとに腫瘍が **局所的** で，注目領域を共有しやすい
  - 一貫性制約が学習を補助する
- Chest X-Ray：
  - **病変の位置・広がりが多様**で，クラス内で注目領域を一意に定めにくい
  - 「同じ場所を見ろ」という制約が必ずしも適合しない

# 5. 結論

# 5.1 まとめ

本研究の貢献：

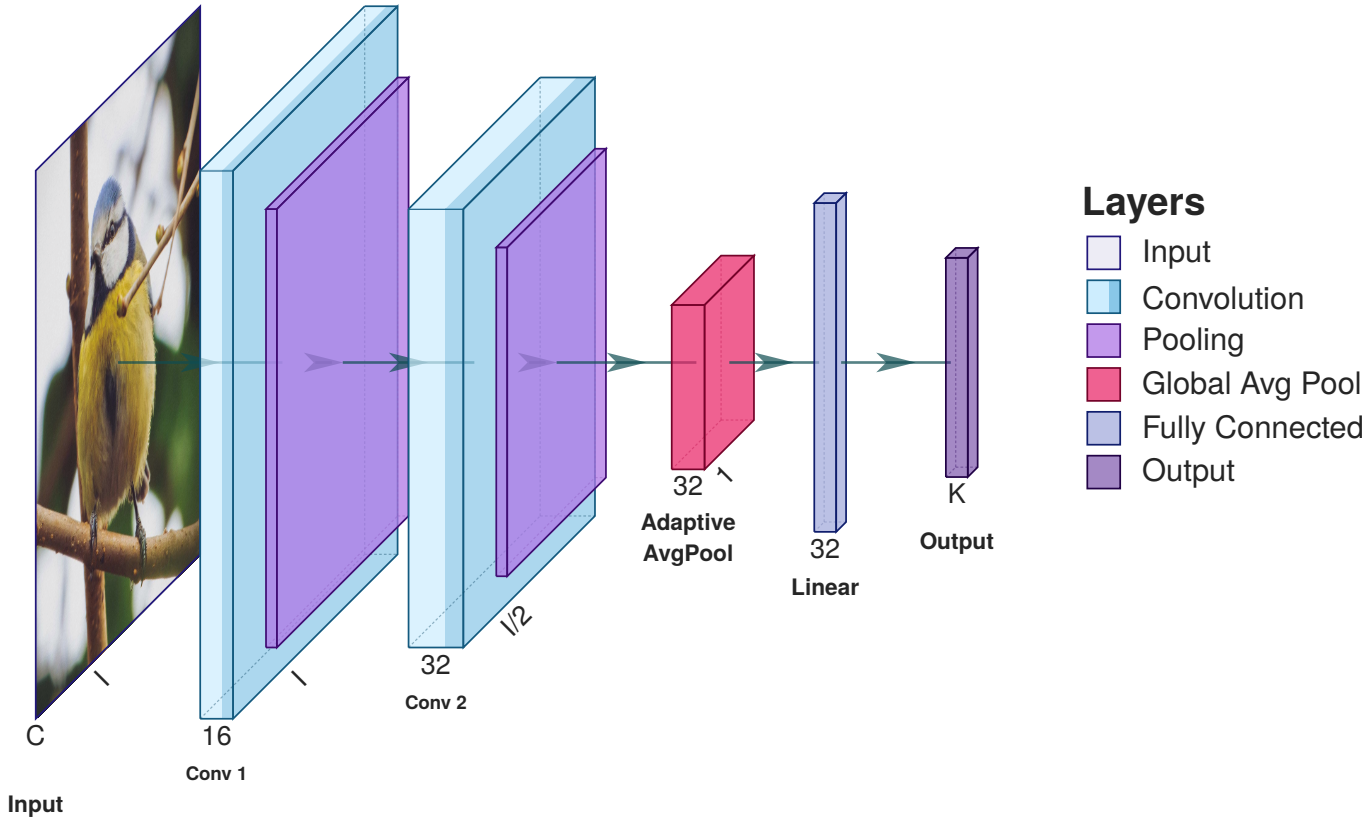
1. Grad-CAM を「**モデルの視線**」とみなして **視線一貫性** $\mathcal{L}_{\mathrm{CAM}}$ を定義
2. **解析的評価** で，性能と一貫性の関係が単純でないことを確認
3. **学習的評価** で，一貫性正則化が特定条件で視線分布を安定化し得ることを確認

- 一貫性の向上が **必ずしも** 性能向上に直結しない（**データセット依存**）
- 一方で，画像構造やタスク特性を見極めて適切に使うなら有用
  → 「**注目領域を狭めるべき**」というタスク特性が明確であれば利用可

# 6. Appendix

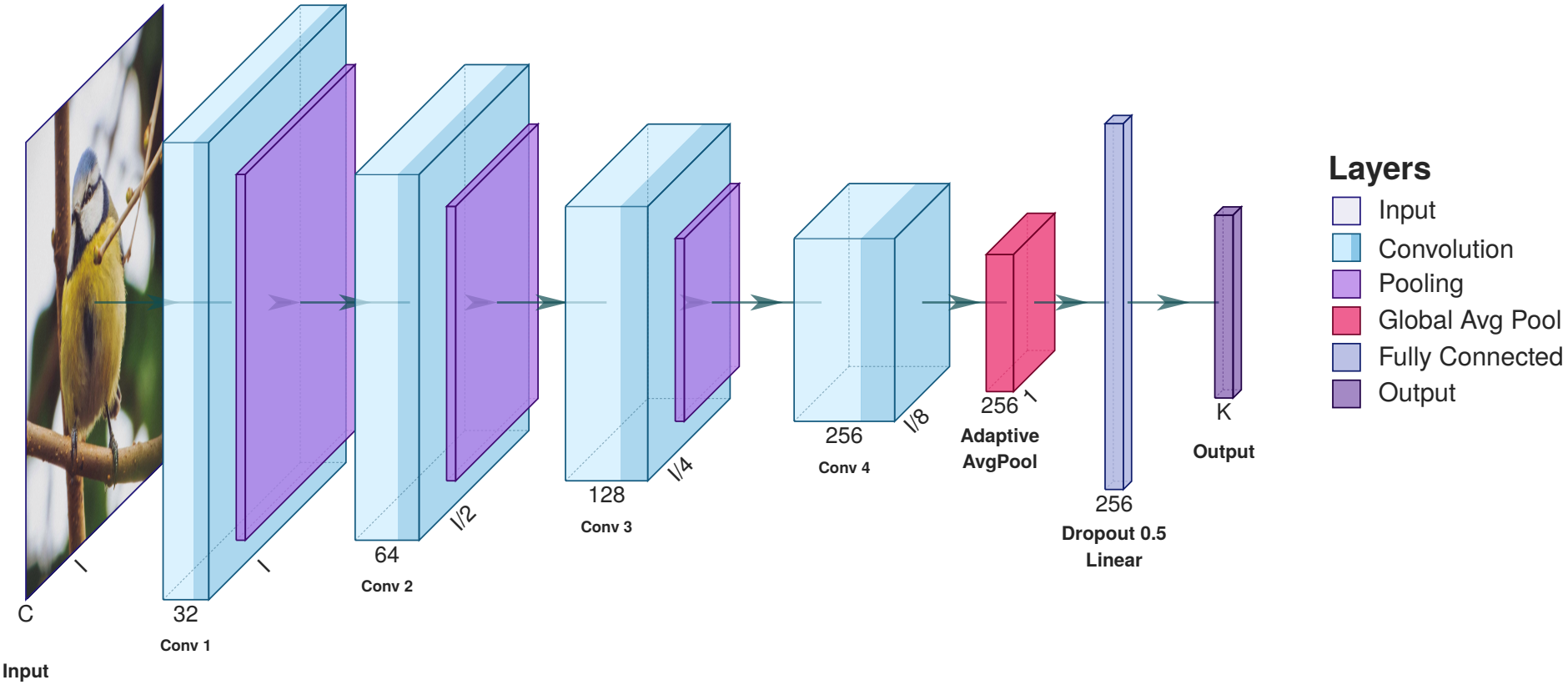
# 各モデルのネットワーク構成図

## Level 1（最小ベースライン）



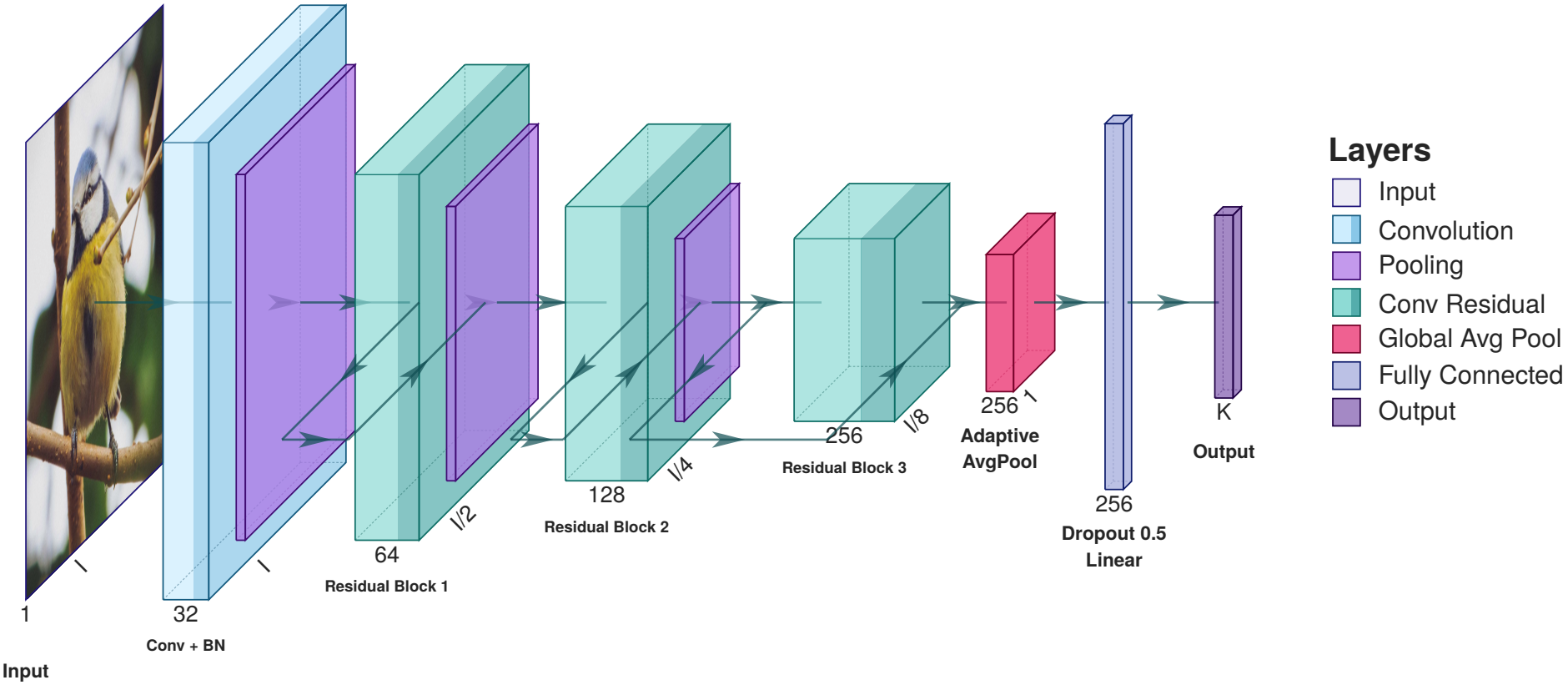
# 各モデルのネットワーク構成図

## Level 2（標準的 CNN）



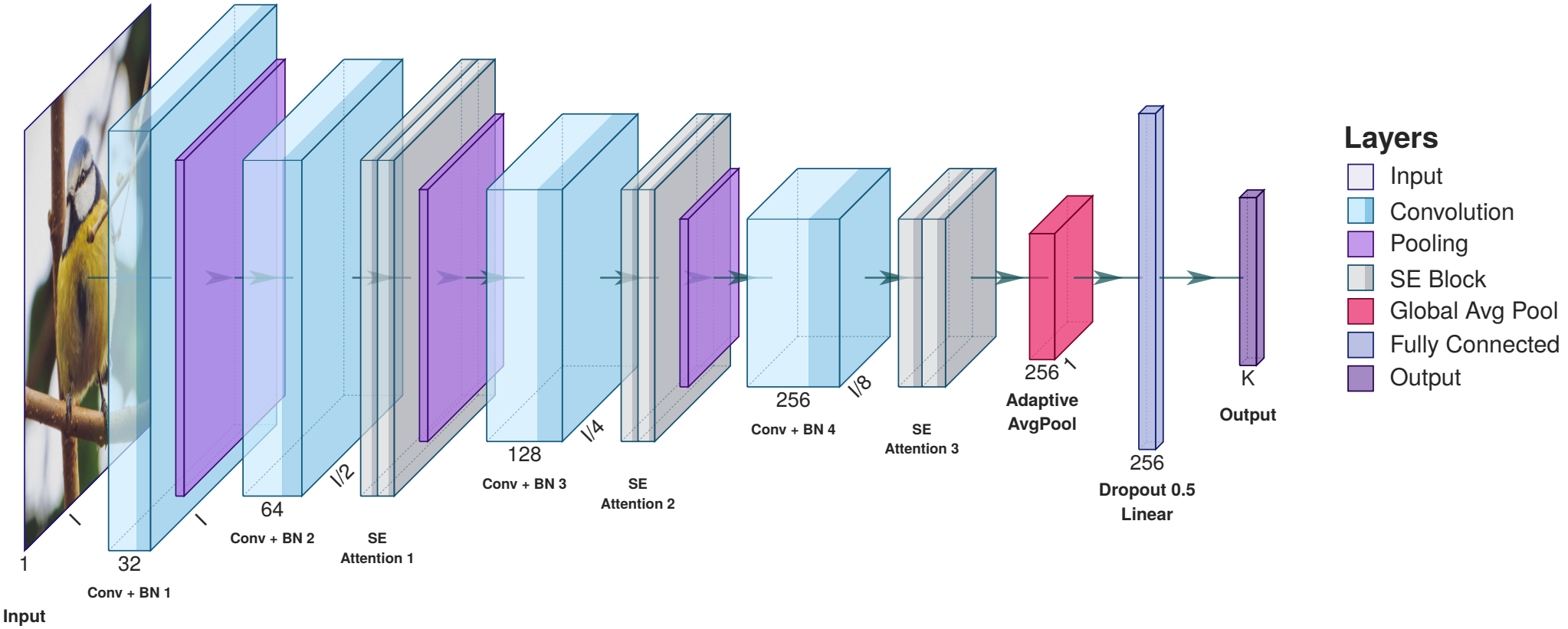
# 各モデルのネットワーク構成図

## Level 3（Residual 構造）



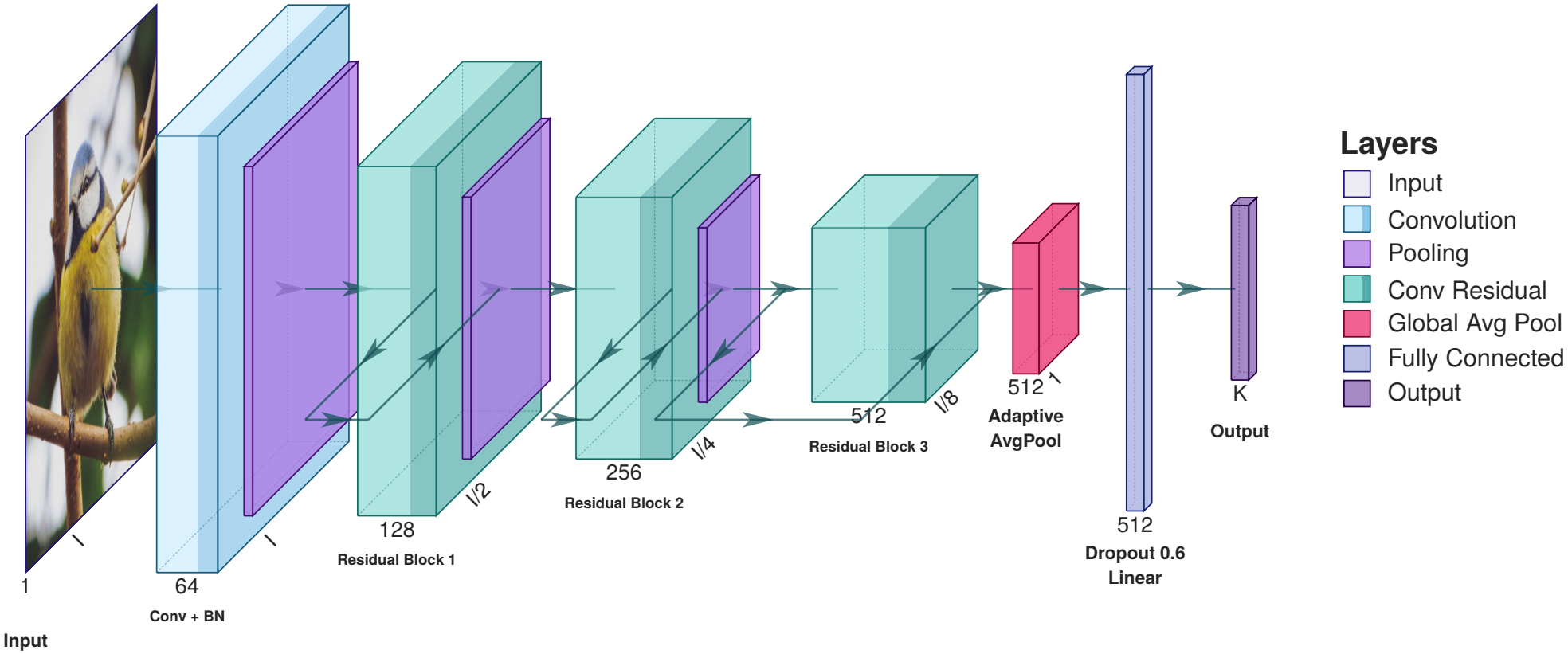
# 各モデルのネットワーク構成図

## Level 4（Attention 機構：SE）



# 各モデルのネットワーク構成図

## Level 5（高表現力）



# Grad-CAM の定義

**Definition 6.1** クラス $c \in C$ に対する Grad-CAM は，対象畳み込み層の特徴マップ $A^k \in \mathbb{R}^{H'W'}(k = 1, ..., M)$ と，クラススコア $y^c \in \mathbb{R}$ の勾配から次で定義される．

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$$

$$\mathrm{CAM}^{(c)} = \mathrm{ReLU}\left( \sum_k \alpha_k^c A^k \right)$$

$Z$:特徴マップの空間サイズに基づく正規化定数　$M$:対象層における特徴マップ数

クラス $y^c$ を大きくする方向に寄与した特徴マップ $A^k$ の勾配を取得

# 評価用の視線一貫性

## Definition 6.1

1. データセット $D = \{(x_i, y_i)\}_{i=1}^{|D|}$ とし，クラス $c$ の正解予測サンプルを

$$D_c^+ := \{(x_i, y_i) \in D \mid \hat{y}_i = y_i = c\}$$

と定義する．→ 正解サンプルのみを抽出

2. クラス $c$ の代表視線（平均 CAM）を

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

と定める．→ 正解サンプルの CAM の平均マップを計算

# 評価用の視線一貫性

3.各サンプルの視線偏差を

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

とし，データセット全体の評価指標を

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

する．→ 各正解サンプルについて，平均マップとの距離（L2 距離）を求め，それを合計・正規化

$\mathcal{L}_{\mathrm{CAM}}$ の **値が小さいほど一貫性が高い** 点に注意

# 正則化済み損失関数

**Definition 6.1** 通常の損失関数 $L_{\mathrm{cls}}$ に 正則化項として 一貫性損失 $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}}$ を加えた損失関数を

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

と定義．ただし，$\lambda$ は一貫性損失の寄与度を表す重み係数．

実装では

- **指数移動平均（EMA）** によるクラス平均視線の平滑化
- 学習初期の不安定性を避ける **カリキュラム学習**

も導入した．

# 学習用の一貫性損失

- 学習時には，データセット単位ではなくミニバッチ単位で評価
- その他，安定した学習のための工夫が含まれる

## Definition 6.1

1. ミニバッチ $B = \{(x_i, y_i)\}_{i=1}^{|B|}$ とし，クラス $c$ の正解予測サンプルを

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

2. クラス $c$ の代表視線（平均 CAM）を

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

3. 各サンプルの視線偏差を

# 学習用の一貫性損失

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

とし，ミニバッチ全体の評価指標を

$$\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}} := \begin{cases} \dfrac{1}{|B^{+\prime}|} \displaystyle\sum_{c \in C'} \sum_{x_i \in B_c^+} \ell_{\mathrm{CAM}}(x_i, c) & (C' \neq \emptyset) \\ 0 & (C' = \emptyset) \end{cases}$$

する．ただし，$C' = \{c \mid |B_c^+| \geq 2\}$，$B^{+\prime} = \bigcup_{c \in C'} B_c^+$

→ 正解サンプルが十分集まらなかったクラスについては計算から除外

万が一，正解サンプルがなかった場合は 0 と定義

# Cohen's κ 係数

- 正答率は **偶然一致** の影響を受ける
  → 今回のような異なるデータセット間での分類性能の比較が難しい

> 2 択のクイズで正答率 50% は「偶然」，4 択のクイズで正答率 50% は「偶然よりも正確」 → 単なる正答率では測れない

- 偶然一致を考慮し $[-1, 1]$ で正規化した「正答率」として利用

**Definition 6.1** κ 係数は観測一致確率 $p_o$ と偶然一致確率 $p_e$ から

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

# なぜ正解予測のみを使う？

視線一貫性は「正しい判断の根拠が揃っているか」を測りたい

- 誤予測の CAM を混ぜると，
  - そもそも「どのクラスの根拠か」が崩れる
  - 誤りの原因（ノイズ・別特徴）を一貫性として数えてしまう
- したがって，
  - 評価用：$D_c^+$（正解予測サンプル）に限定
  - 学習用：$B_c^{+\prime}$ （$|B_c^+| \geq 2$ を満たすクラスの正解サンプル）のみで損失計算

# なぜ L2 距離を使う？

CAM を $[0, 1]$ に正規化しているため，**画素ごとの差** を素直に測れる距離が扱いやすい

- L2（二乗誤差）：
  - 微分可能で安定した勾配が得やすい
  - 平均との差（分散的量）として解釈しやすい
  - ただし，指標値が $10^{-3}$ に集中しやすいなど，**スケールの解釈性** は今後の改善点
- 他の距離指標の候補: **L1 距離**，**cosine 距離**，**SSIM**，**JS ダイバージェンス**，**Earth Mover's Distance** …

# なぜ Level 1 モデルで正の相関が出た？

詳しくはまだ分かっていない． ただし，仮説はある．

- Level 1 については，相関係数以外でも **分類性能** $\kappa$ や **視線一貫性** $\mathcal{L}_{\mathrm{CAM}}$ の分布がほかモデルと大きく異なっていた
  → Level 1 は相関以外でも**特異性**があった
- Level 1 は比較実験のために用意した **非常に表現力の小さいモデル**（現実的・実用的なラインよりも下の表現力）
  → モデルが少ない表現力の中で，正確に分類するために様々なところを見ようとした結果，視線が逆に広がった