令和７（2025）年度

理学部　卒業論文

# 卒業論文タイトル

Vision Transformer との比較に基づく畳み込みニューラルネットワークのマルチモーダル情報統合における阻害要因の解析

分 野 名　　物質科学

学籍番号　　222045

氏　　名　　木村裕健

指導教員　　ミケレットルジェロ

令和８年１月　日　提出

# 目次

# 1. 序論

## 1.1 背景

人間の脳は、外界から入力される視覚、聴覚、言語といった多様な感覚情報を統合し、意味のある知覚を形成する高度な情報処理機関である 。知覚情報科学において、この複雑な情報処理機構を計算論的に明らかにすることは、人間の知覚メカニズムの理解のみならず、人工知能（AI）のさらなる発展においても極めて重要な課題である。特に人間の視覚系は、網膜が捉えた物理的な画素データをそのまま処理するのではなく、脳内での予測や先行知識、および随伴する文脈情報を統合することで、対象を瞬時に認識・分類している（Eagleman 2001）。

現代のデジタル社会において、TikTok に代表されるショート動画プラットフォームは、視覚情報（映像）、色彩的特徴、および言語情報（キャプションやハッシュタグ）が高度に絡み合う「マルチモーダルな情報空間」を形成している。人間がこれらの膨大なコンテンツを「ダンス」「グルメ」「旅行」といったカテゴリに瞬時に分類する際、脳内では単なる形状認識を超え、色彩による心理的影響や文字情報の文脈、さらにはエンゲージメント（いいね数や保存数）といった社会的・統計的情報を統合的に処理していると考えられる。

近年の神経科学や知覚科学では、脳の理論的・計算論的なモデルを構築し、その振る舞いを調べることで知覚メカニズムを理解する手法が重要視されている。AI 分野における畳み込みニューラルネットワーク（CNN）は、脳の一次視覚野（V1）における局所的な特徴抽出から、高次視覚野へ至る階層的な情報処理機構を模倣することで発展してきた。一方、近年登場した Vision Transformer（ViT）は、CNN のような生物学的な構造制約に縛られず、画像内のパッチ間の広域的な依存関係（注意機構）を直接捉える特性を持つ。これらの異なる計算論的モデルを用いて SNS コンテンツの分類タスクを行うことは、工学的な精度向上のみならず、マルチモーダルな情報統合がいかにして知覚を形成するかを検証する上で極めて重要な意義を持つ。

## 1.2 目的

本研究では、計算論的なアプローチから、SNS 動画のジャンル分類における視覚情報とメタデータの統合メカニズムを検証する。具体的には、独自に取得した 3 つのクラス（「ダンス」「グルメ」「旅行」）のデータセットを用い、局所的な特徴抽出に長けた CNN（EfficientNet-B0）と、全体的な文脈把握に長けた ViT の二つのモデルの振る舞いを比較する。

さらに、画像情報に加えて「社会的知覚（インプレッション）」「色彩的特徴（HSV/RGB 統計量）」「被写体属性（顔の数）」「言語的特徴（文字情報）」の計 16 次元のメタデータを統合する。この際、情報の統合手法として画像バックボーンから抽出された特徴ベクトルとメタデータの特徴量を結合する「後方融合（Late Fusion / 特徴レベル統合）」を採用する。これは、脳が高次領域で異種情報を統合するプロセスを模倣したものである。

本研究を通して、CNN と ViT の構造的違いが分類精度や判断根拠に与える影響を明らかにするとともに、マルチモーダルな情報の付与が「AI の知覚」をいかに人間に近づけ得るかを、知覚情報科学の観点から考察することを目的とする。

# 2．研究手法

## 2.1 CNN (畳み込みニューラルネットワーク)

CNN (畳み込みニューラルネットワーク)とは、入力画像から局所的な特徴を抽出するための主要なコンポーネントである。LeCun ら（1998）が提案したモデル（LeNet-5）において、第$l$層の特定の特徴マップにおける座標$(i,j)$の値 $y_{ij}^{l}$ は、前層（第$l-1$層）の出力とフィルタ（カーネル）との離散畳み込み演算によって算出される。一般的な離散畳み込み演算の形式は、以下の式で定義される。

$$y_{i,j}^{(l)}=\sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}w_{m,n}^{(l)}\cdot x_{i+m,j+n}^{(l-1)}+b^{(l)}\right)\#(2.1)$$

ここで、各変数は以下の通りである：

- $x_{i+m,j+n}^{(l-1)}$：前層（第$l-1$層）における入力データの局所領域の値
- $w_{m,n}^{(l)}$：$M\times N$ サイズを持つフィルタ（カーネル）の重み係数
- $b^{(l)}$：第$l$層のバイアス項
- $\sigma$：活性化関数（当時はシグモイド関数や tanh が主流であったが、現在は$ReLU$が一般的である）

この演算により、フィルタが画像全体を走査（スライド）しながら、エッジや特定の形状パターンとの適合度を計算し、特徴マップとして出力する。この際、同一のフィルタ重みを画像全体で使い回す「重み共有（Weight Sharing）」の仕組みにより、学習すべきパラメータ数を劇的に削減している。また、この構造的特性により、画像内における物体の

位置が多少変化しても正しく特徴を捉えることができる「平行移動不変性（Translation Invariance）」を実現している。

## 2.2 ViT Attention 機構とは

Transformer の核となるのは、Scaled Dot-Product Attention と呼ばれる機構である。これは、入力データの中から「どこに注目すべきか」を動的に決定する仕組みである。
Vaswani ら（2017）によれば、Attention 層への入力は Query（Q）、Key（K）、Value（V） の 3 つから構成される。出力は Value の加重平均として計算され、その重みは Query と対応する Key の適合度（類似度）によって決定される。具体的な計算式は以下の通りである。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \#(2.2)$$

ここで、各変数は以下の通りである：

- $Q$(Query)：検索をかける側のベクトル（問い合わせ）。現在の注目地点を表す。
- $K$(Key)：検索対象のインデックスとなるベクトル。 との適合度を計算するために用いられる。
- $V$(Value)：抽出される情報の本体となるベクトル。
- $d_k$：ベクトルの次元数。
- $\frac{1}{\sqrt{(d_k)}}$：スケーリング因子。次元数が大きくなった際に、内積の値が過大になり勾配が消失するのを防ぐ役割を持つ。
- $softmax$：内積の結果を正規化し、合計が 1 になるような重み（Attention Weight）に変換する関数。

## 2.3 評価指標

評価指標:正解率（Accuracy）
本研究では、提案モデルの画像分類性能を定量的に評価するための主要な指標として、正解率（Accuracy）を採用する。正解率は、全テストサンプル数に対して、モデルが正しく予測したサンプルの割合を示す。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \#(2.3)$$

画像認識分野において、Krizhevsky ら (2012) は大規模画像データセット（ImageNet）を用いたコンテストにおいて深層畳み込みニューラルネットワーク（CNN）の有効性を証明し、その評価において誤差率（Error Rate）を用いた比較検討を行った。現在、多くの画像分類タスクにおいては、Krizhevsky らの手法を継承し、その補数である正解率がモデルの総合的な識別能力を測る標準的な指標として広く用いられている。本実験においても、先行研究との比較可能性を確保するため、この指標を用いて評価を行う。

評価指標：F1-score（F 値）
本研究におけるモデルの性能評価には、精度の偏りを総合的に評価するため、F1-score（F 値）も採用する。F1-score は、適合率（Precision）と再現率（Recall）の調和平均によって算出される指標である。適合率は「モデルが正と予測したもののうち、実際に正であった割合」を示し、再現率は「実際に正であるもののうち、モデルが正と予測できた割合」を示す。
一般に適合率と再現率はトレードオフの関係にある。単純な正解率（Accuracy）のみでは、データセットのクラス分布に偏りがある場合（不均衡データ）、多数派のクラスを予測するだけで高い値が出てしまい、モデルの真の識別性能を正しく評価できない恐れがある。F1-score を用いることで、これら 2 つの指標を等しく考慮し、モデルの汎用性をより厳密に評価することが可能となる。
F1-score の算出式を以下に示す。

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \#(2.4)$$

ここで、各指標は混同行列における真陽性（True Positive: TP）、偽陽性（False Positive: FP）、偽陰性（False Negative: FN）を用いて次のように定義される。

$$Precision = \frac{TP}{TP + FP} \#(2.5)$$

$$Recall = \frac{TP}{TP + FN} \#(2.6)$$

# 3. 理論的背景・関連研究

本章では、本研究の核となる画像分類モデルの進化、特に畳み込みニューラルネットワーク（CNN）と Vision Transformer（ViT）の構造的差異と、それがもたらす帰納バイアス（Inductive Bias）の違いについて詳述する。また、近年のマルチモーダル学習において課題となっている「モダリティ間の競合（Modality Competition）」や「Fusion（統合）」に関する理論的背景を整理し、本研究の実験設定である「バックボーン凍結

（Frozen Backbone）」および「Late Fusion」がモデルの挙動に与える影響を考察するための基盤を構築する。

## 3.1 視覚的特徴抽出の変遷と帰納バイアス

### 3.1.1 CNN における局所受容野と平行移動不変性

CNN は、画像処理に適した強力な帰納バイアス（Inductive Bias）を有しており、それは主に以下の 2 点に集約され 。

1. 局所性（Locality）: フィルタのサイズ（例：$3 \times 3$）に基づき、各ニューロンは画像の限定的な領域（局所受容野）のみを参照する。これにより、エッジやテクスチャといった局所的な特徴抽出が可能となる。
2. 重み共有（Weight Sharing）: 同一のフィルタを画像全体にスライドさせて適用することで、対象物の位置によらず特徴を抽出できる「平行移動不変性（Translation Invariance）」を獲得する。

しかし、これらの制約は大域的な情報の統合を困難にする要因ともなり得る 。画像全体の関係性（例：背景の空と手前の被写体との位置関係）を理解するためには、プーリング層によるダウンサンプリングを伴う多層構造によって受容野を段階的に広げる必要がある。この構造的特性は、ネットワークの低層部において「意味（Semantics）」よりも「テクスチャ（Texture）」が支配的になる傾向を示唆している。

### 3.1.2 Transformer と Attention 機構によるパラダイムシフト

一方、自然言語処理（NLP）の分野では、Vaswani ら（2017）による「Attention Is All You Need」の提案により、革新的な転換が起きた 。彼らが提案した Transformer アーキテクチャは、再帰型ニューラルネットワーク（RNN）や CNN が持つ「近傍の要素から順に処理する」という制約を取り払い、Self-Attention 機構によってシーケンス内の全要素間の関係性を直接計算することを可能にした。

前述の数式(2)において、Query ($Q$)、Key ($K$)、Value ($V$) は入力から生成されるベクトルであり、$\frac{Q K^T}{\sqrt{(d_k)}}$ は各要素間の類似度（注意の重み）を表す。この計算には「距離」の概念が含まれていないため、シーケンスの先頭にある単語と末尾にある単語であっても、CNN のように層を深くすることなく、第 1 層から直接相互作用（Global Context の獲得）が可能となる。この「距離に依存しない大域的な依存関係の抽出能力」は、画像の文脈や構図を捉える上で極めて有利な特性であり、Vision Transformer（ViT）の性能を支える基盤となっている。

## 3.2 画像認識におけるテクスチャバイアスと形状バイアス

### 3.2.1 ImageNet で学習された CNN のテクスチャ偏重

本研究で用いる EfficientNet を含む CNN モデルの多くは、大規模画像データセット ImageNet（Krizhevsky et al., 2012）で事前学習されている 。しかし、Geirhos ら（2019）は、CNN の認識メカニズムに関する重要な知見を提示した。彼らは「ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness」と題した論文において、CNN が人間のように物体の「形状（Shape）」を見て判断しているのではなく、局所的な「テクスチャ（Texture）」に過度に依存していることを明らかにした。

Geirhos らは、スタイル変換（Style Transfer）を用いて、物体の形状とテクスチャを矛盾させた「Cue-Conflict 画像」を作成した（例：象の皮膚のテクスチャを持つ猫の画像）。実験の結果、人間はこれを「猫」と認識（形状バイアス）したのに対し、ResNet-50 などの標準的な CNN は圧倒的に「象」と誤分類（テクスチャバイアス）したのである。

この事実は、本研究における「バックボーン凍結」の設定に対し、重要な示唆を与える。CNN バックボーンを凍結して使用する場合、その出力は事前学習時のテクスチャ情報が支配的な表現となる可能性が高い 。ソーシャルメディアにおける「ダンス」や「旅行」といったカテゴリは、個別のテクスチャよりも被写体のポーズや風景の構成といった大域的な情報に依存するため、凍結された CNN 特徴量ではタスクに不可欠な意味情報を十分に捉えきれない懸念がある。

### 3.2.2 Vision Transformer における形状バイアスの獲得

Dosovitskiy ら（2020）は、Transformer アーキテクチャを画像認識に適用した Vision Transformer（ViT）を提案した 。ViT は画像を$16 \times 16$ピクセル等の固定サイズのパッチに分割し、それらを線形投影（Linear Projection）によって埋め込みベクトル化することで、シーケンスデータとして処理する。

Dosovitskiy らは、ViT が CNN とは対照的に、画像特有の帰納バイアス（局所性や平行移動不変性）をあえて排除した設計であることを指摘している 。その代わり、ViT は Self-Attention 機構を通じて、初期層から画像全体にわたる大域的な情報を統合的に学習する 。この特性により、ViT は CNN と比較して「形状バイアス（Shape Bias）」が強く、人間の視覚認識に近い大域的な構造把握を行うことが報告されている 。また、最終層の分類トークン（[CLS] token）には画像全体のパッチ情報が高度に集約されており、その表現空間は個々のピクセル値の依存から脱却した、抽象度の高い「意味的（Semantic）」な特徴を保持していると考えられる 。

## 3.3 マルチモーダル学習と Fusion の課題

### 3.3.1 マルチモーダル融合手法の分類

画像、テキスト、数値データなどの異種情報（マルチモーダルデータ）を統合する手法は、統合のタイミングに基づき主に以下の 3 つに分類される。

1. 前方融合（Early Fusion / 特徴レベル融合）各モダリティから特徴を抽出する初期段階、あるいは生の入力データの段階で結合を行う手法である。モダリティ間の相関関係を低次レイヤーから直接学習できる利点がある一方、各データの時間的・空間的な整列（Alignment）が不可欠であり、入力次元の増大に伴う計算コストの増加を招きやすい。
2. 後方融合（Late Fusion / 決定・特徴レベル融合） 各モダリティを独立したバックボーンネットワークで処理し、最終的な出力（ロジット）あるいは抽出された特徴ベクトルを結合する手法である。本研究では、画像バックボーンより得られた視覚特徴と、多層パーセプトロン（MLP）により変換されたメタデータ特徴を、全結合層の直前で連結する「特徴レベルの後方融合」を採用する。本方式は各モダリティの独立性が高く、事前学習済みモデルの転用や特定のモダリティの差し替えが容易であるという高い汎用性を持つ 。
3. 中間融合（Intermediate Fusion / 融合ネットワーク） ネットワークの中間層において、Cross-Attention 機構などを用いてモダリティ間の相互作用を逐次的に学習させる手法である。特定の階層で情報の交換を行うことで、より高度な意味的関連性を抽出することが可能となるが、アーキテクチャ設計の複雑性が増す傾向にある。

### 3.3.2 モダリティの競合（Modality Competition）

近年、Huang ら（2022）は、マルチモーダル学習における「Modality Competition（モダリティの競合）」という現象を理論的に解明した。彼らの指摘によれば、複数のモダリティを同時に学習（Joint Training）させる際、特定のモダリティ（一般に学習が容易な情報源）が最適化を支配し、他のモダリティの学習が抑制される現象が生じる。

特に、本研究のように「バックボーンを凍結（Frozen Backbone）」する設定下では、この問題が顕在化しやすい。画像エンコーダの重みは更新されず、結合後の全結合層のみが学習されるため、画像特徴量がタスクに対して不十分（ノイズの混入やテクスチャへの過度な偏り）である場合、モデルは画像情報を事実上無視し、学習可能なパラメータを持つメタデータ側の入力に過剰適合（Overfitting）する傾向が強まる。

Wu ら（2022）や Wang ら（2024）は、こうした勾配の干渉を抑制するためにモダリティの分離（Decoupling）手法を提案しているが、単純な連結（Concatenation）ベースの手法においては、勾配が流れやすいモダリティが静的なモダリティ（画像）の寄与を減

退させてしまうリスクが依然として残る 。このメカニズムこそが、本研究の実験結果における精度の逆転現象を考察する上での重要な鍵となる。

## 3.4 モデルの諸元と規模の比較

本研究で比較対象とする二つのアーキテクチャは、その設計思想のみならず、モデル規模（表現能力の容量）においても大きな開きがある。

- EfficientNet-B0: パラメータ数は約 530 万であり、モバイル端末等での効率性を重視した軽量モデルである。
- ViT-B/16: パラメータ数は約 8,600 万であり、EfficientNet-B0 の約 16 倍以上の容量を有する。

この規模の差は、実験 2 におけるマルチモーダル情報の受容能力や、ノイズに対する耐性に影響を与える重要な前提条件となる。

# 4. 画像データセットの構築とメタデータ解析

## 4.1 データセットの構築

本研究では、ショート動画プラットフォーム「TikTok」から取得した画像を対象にデータセットを構築した。

4.1.1 画像の取得方法

画像データは、TikTok の検索機能を用い、「ダンス」「旅行」「グルメ」の 3 つのジャンルを対象に収集した。取得の詳細は以下の通りである。

- 取得日時: 2025 年 8 月 23 日（ダンス、旅行）、2025 年 8 月 25 日（グルメ）
- 取得環境: Mac M4 環境にてスクリーンショットとして保存
- データ量: 各ジャンル 100 枚、合計 300 枚

4.1.2 分析環境

収集したデータの分析および特徴抽出には、Google Colab Pro（GPU: NVIDIA Tesla T4 / A100）を使用した。使用した主なライブラリは、特徴抽出に PyTorch および TensorFlow、データ処理に Pandas、可視化に Seaborn および Matplotlib である。

## 4.2 メタデータの定義と入力次元数

本研究では、コンテンツの視覚的特徴、社会的反応、および意味的コンテキストを定量化するため、実験フェーズごとに最適化したメタデータベクトルを構築した 。各実験で使用したメタデータの具体的な構成と次元数を表 4.1 に示す。

表 4.1：メタデータの定義と抽出

| カテゴリ | 具体的な変数 | 実験 1 | 実験 2 |
|---|---|---|---|
| HSV 色空間値 | mean_hue, mean_saturation, mean_value | 3 | 3 |
| RGB 統計量 | mean, std, skew, kurtosis（R, G, B） | 12 | 6 |
| 社会的知覚（Impression） | likes, saves, comments | 1（likes のみ） | 3 |
| 被写体属性（Face） | face_count | 0 | 1 |
| 言語的特徴（Text） | text_len, has_text, keyword_count | 0 | 3 |
| 合計次元数 | | 16 | 16 |

4.2.1 画像統計量の詳細

画像統計量は、単一のファイル（hsvdata.csv）から取得されているが、実験の目的に応じて以下の通り選択する変数を変更している 。

実験 1 の設計: 画像自体の統計的性質を重視し、hsvdata.csv に含まれる全ての数値データ（15 次元）に、インプレッションの 1 次元（likes）を加えた合計 16 次元のベクトルを採用した。ここには各色の平均だけでなく、歪度（skew）や尖度（kurtosis）といった高次の統計量が含まれる 。

実験 2 の設計: 画像統計量を HSV の 3 次元および RGB の平均・標準偏差の 6 次元（計 9 次元）に厳選し、空いた次元に「顔の数」や「テキスト情報」といった意味的なコンテキストを補完した 。これにより、合計次元数を 16 に維持したまま、マルチモーダルな情報の多様性を向上させている 。

## 4.3 メタデータの統計的特性

抽出したメタデータの基本統計量を表 4.2 に示す。

表 4.2：メタデータの基本統計量

| 変数名 | 平均値 (μ) | 標準偏差 (σ) | 最小値 | 中央値 | 最大値 |
|---|---|---|---|---|---|
| いいね数 | 197,226 | 1,090,224 | 0 | 29,500 | 13,000,000 |
| 保存数 | 13,031 | 27,266 | 0 | 3,851 | 305,300 |
| コメント数 | 3,819 | 37,351 | 0 | 195 | 510,000 |
| 平均色相 (Hue) | 0.32 | 0.17 | 0.00 | 0.29 | 0.93 |
| 平均彩度 (Sat) | 0.30 | 0.14 | 0.01 | 0.29 | 0.81 |
| 平均明度 (Val) | 0.51 | 0.15 | 0.00 | 0.53 | 0.91 |
| 検出顔数 | 0.52 | 0.95 | 0 | 0 | 7 |

## 4.4 メタデータ分析

各変数の相関およびカテゴリ別の特徴について分析を行った。

### 4.4.1 エンゲージメントの相関分析

変数間の相関をヒートマップ（図 4.1）および散布図（図 4.2）で確認したところ、保存数（Saves）といいね数（Likes）の間に強い正の相関（r=0.71）が認められた 。

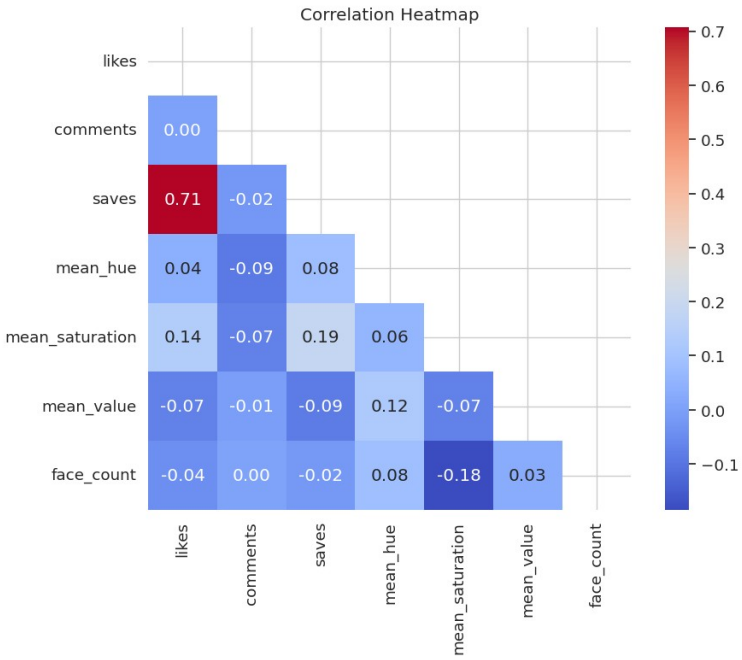



図 4.1：変数間の相関ヒートマップ

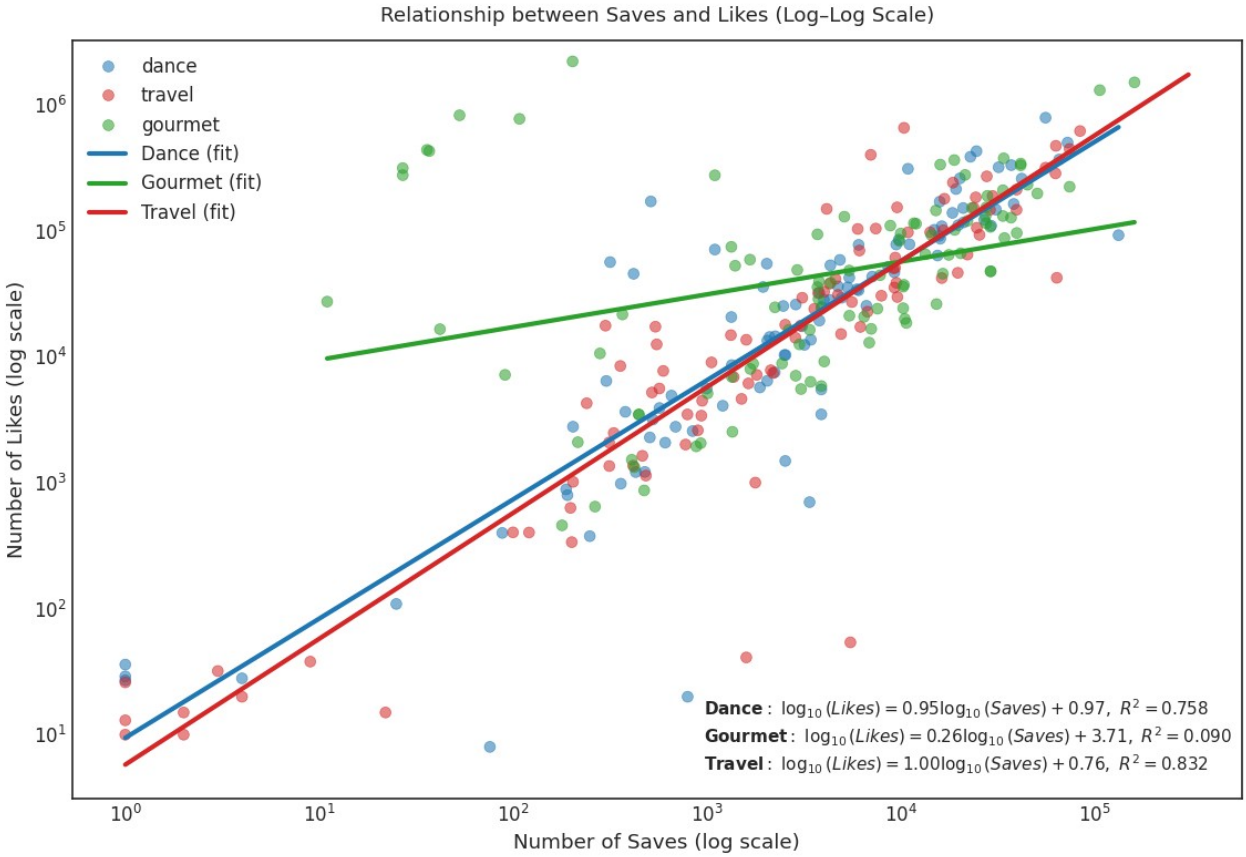



図 4.2：保存数といいね数の相関関係（対数スケール）

ジャンル別の回帰分析では、以下の結果が得られた 。

- Dance: $log_{10}(Likes) = 0.95\ log_{10}(Saves) + 0.97,\ R^2 = 0.758$
- Travel: $log_{10}(Likes) = 1.00\ log_{10}(Saves) + 0.76,\ R^2 = 0.832$
- Gourmet: $log_{10}(Likes) = 0.26\ log_{10}(Saves) + 3.71,\ R^2 = 0.090$

「Travel」と「Dance」は高い相関を示したが、「Gourmet」カテゴリのみ相関が極めて低く、独自のエンゲージメント傾向を持つことが示唆された。

### 4.4.2 カテゴリ別特徴の比較

カテゴリ別の「いいね数」の平均および標準偏差を表 4.3 および図 4.3 に示す。

表 4.3：カテゴリ別エンゲージメント統計（いいね数）

| カテゴリ | 平均値 (Mean) | 標準偏差 (STD) | サンプル数 |
|---|---|---|---|
| Travel | 330,047.60 | 1,823,443 | 100 |
| Gourmet | 186,031.76 | 464,909 | 100 |
| Dance | 75,598.40 | 125,719 | 100 |

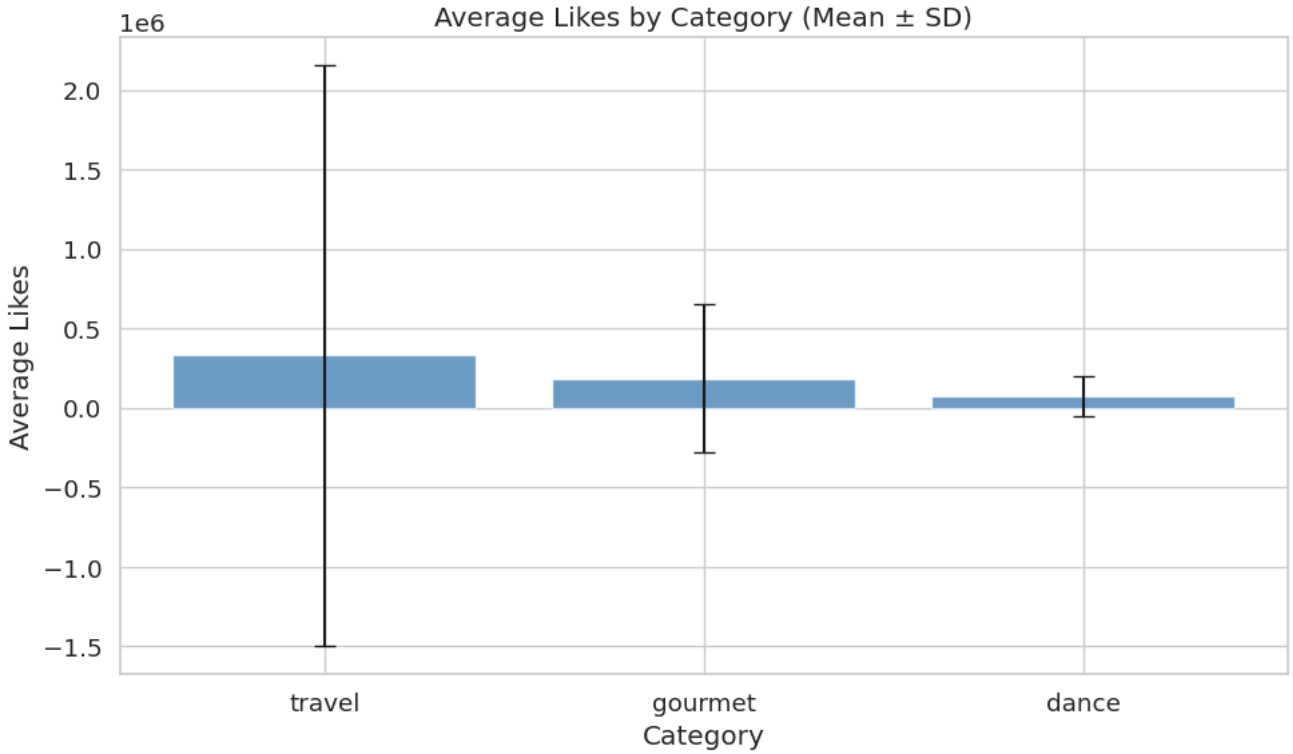



図 4.3：カテゴリ別の平均いいね数

次に、画像特性としての HSV 色空間および顔検出数の統計量を表 4.4 および表 4.5 に示す。

表 4.4：カテゴリ別 HSV 色空間の統計量（平均 ± 標準偏差）

| カテゴリ | 平均色相 (Hue) | 平均彩度 (Sat) | 平均明度 (Val) |
|---|---|---|---|
| Dance | 0.35 ± 0.20 | 0.23 ± 0.11 | 0.51 ± 0.16 |
| Travel | 0.36 ± 0.16 | 0.36 ± 0.16 | 0.51 ± 0.17 |
| Gourmet | 0.23 ± 0.12 | 0.29 ± 0.09 | 0.49 ± 0.10 |

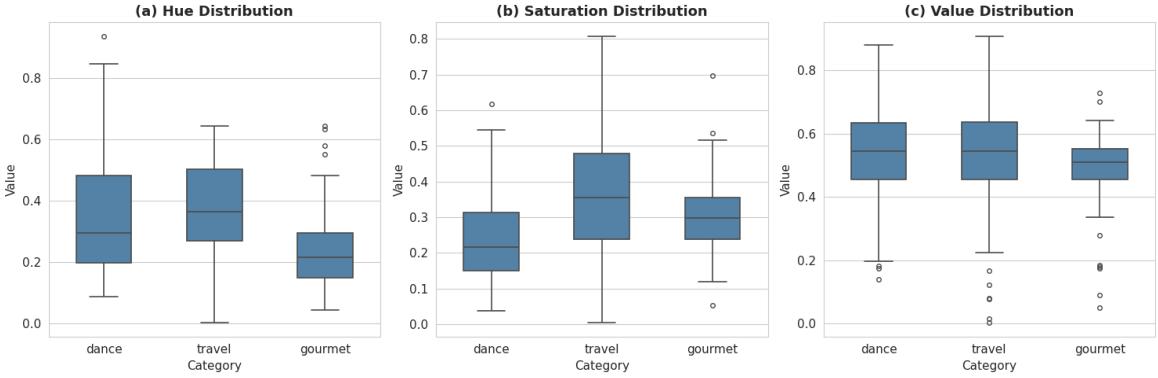



図 4.4：カテゴリ別の HSV 色空間分布

表 4.5：カテゴリ別平均検出顔数

| カテゴリ | 平均値 (Mean) | 標準偏差 (STD) |
|---|---|---|
| Dance | 1.18 | 1.02 |
| Travel | 0.25 | 0.91 |
| Gourmet | 0.14 | 0.40 |

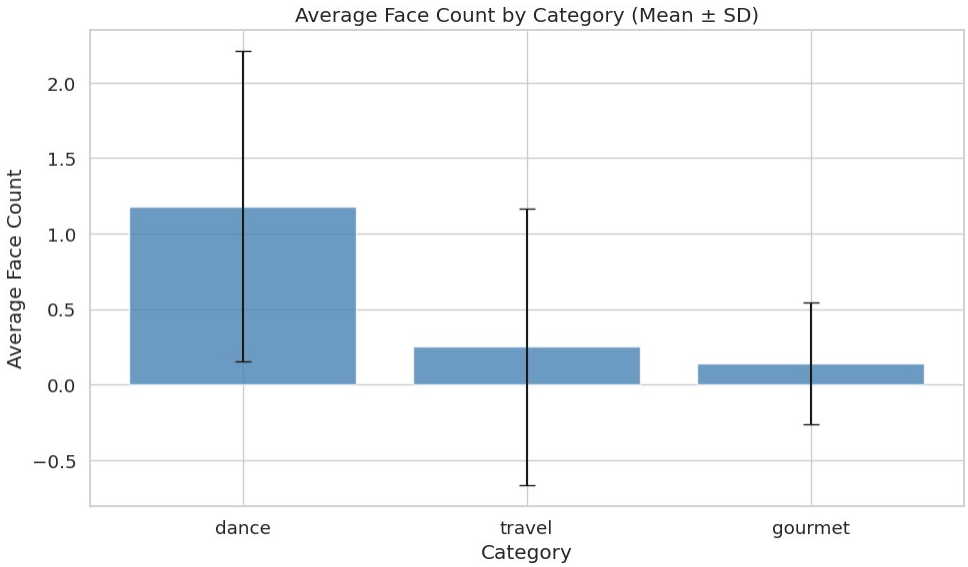



図 4.5：カテゴリ別の平均検出顔数

分析の結果、「Dance」カテゴリは顔検出数が突出して多く、人物がコンテンツの主体であることを示している。一方、「Travel」は彩度が他カテゴリより高く、視覚的に鮮やかな画像が多い傾向が確認された。

### 4.4.3 テキスト情報の特性

OCR によって抽出されたテキストデータの統計的要約を表 4.6 に示す。また、各カテゴリにおける頻出キーワードの分析結果を表 4.7 に示す。

表 4.6：テキスト特性の統計的要約

| カテゴリ | 文字数 (text_len) | 単語数 (word_count) | ハッシュタグ数 |
|---|---|---|---|
| Dance | 14.1 ± 11.6 | 1.8 ± 1.9 | 0.0 |
| Travel | 32.8 ± 65.3 | 4.6 ± 11.3 | 0.01 |
| Gourmet | 17.9 ± 12.1 | 2.8 ± 1.6 | 0.01 |

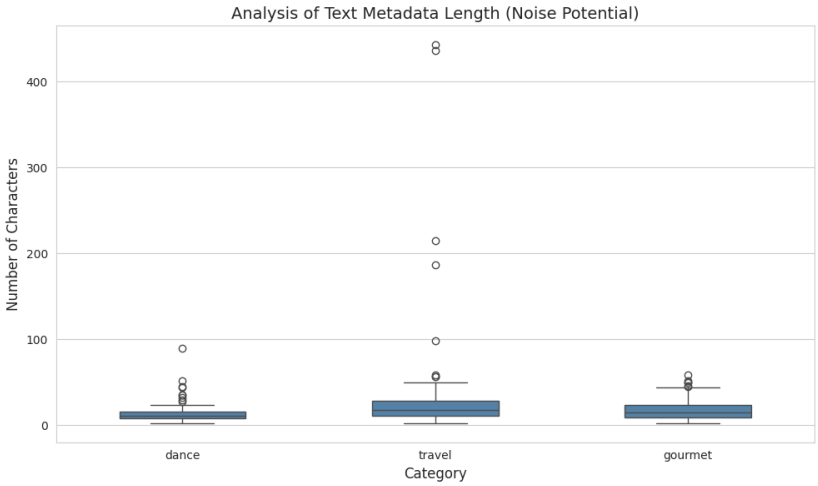



図 4.6：テキスト情報の文字数分布

表 4.7：カテゴリ別頻出キーワード（Top 5）

| 順位 | Dance | Travel | Gourmet |
|---|---|---|---|
| 1 | 関連するコンテンツを見つける(13) | japan(8) | 東京(13) |
| 2 | dance(5) | 1 位(8) | 渋谷(5) |
| 3 | ダンス(4) | 観光(6) | 焼肉(4) |
| 4 | tiktok(4) | 関連するコンテンツを見つける(6) | ラーメン(4) |
| 5 | ダンス女子(3) | 旅行(5) | 銀座(3) |

※ 出現回数を括弧内に記載（例：東京(13)）

## 4.5 まとめ

本章では、TikTok より収集した 3 つのジャンルの画像データセットについて、メタデータを用いた多角的な分析を行った。分析の結果、以下の知見が得られた。

1. エンゲージメントの乖離: Travel および Dance では保存数といいね数に強い相関が見られたが、Gourmet においてはその相関が極めて低く、ユーザーの保存動機が他ジャンルと異なる可能性が示唆された。
2. 視覚的・構造的特徴: Dance は顔検出数が多く、Travel は色彩の鮮やかさ（彩度）に特徴があることが定量的に示された。
3. 言語的特徴: テキストマイニングの結果、各ジャンルの特性を反映した固有のキーワード群が特定された。

# 5. 実験 1：CNN におけるメタデータの統合評価

## 5.1 実験設定

本実験では、画像情報のみを用いる「CNN single」と、画像統計量（HSV および RGB 統計量）15 次元と、インプレッション（likes）1 次元を組み合わせた、計 16 次元のメタデータベクトルを統合した「CNN hybrid」の比較を行った。

本実験における「CNN hybrid」の特徴は、バックボーンネットワーク（EfficientNet-B0）を凍結（Frozen）した状態で学習を行った点にある。CNN を固定された特徴抽出器として運用したことで、統計情報がノイズとならずに画像を補完した。

統合したメタデータは、表 4.1 に定義した項目のうち、以下の 2 種類である。

- Impression：いいね数に基づく社会的知覚情報
- HSV 色空間値：画像統計量：画像の色相・彩度・明度（HSV）および RGB の統計値に基づく色彩的特徴量

これらの抽象度が低く、かつ画像特徴と親和性の高い統計情報を付加することで、識別精度への影響を検証した。

## 5.2 実験結果

5 回の試行における正解率（Accuracy）および損失関数（Loss）の平均値を表 5.1 と図 5.1 に示す。

表 5.1：実験 1 における正解率と損失関数の平均・標準偏差

| | Mean Accuracy | Std Dev Accuracy | Mean Loss | Std Dev Loss |
|---|---|---|---|---|
| CNN single | 0.8633 | 0.0287 | 0.3441 | 0.0421 |
| CNN hybrid | 0.9067 | 0.0271 | 0.2400 | 0.0277 |

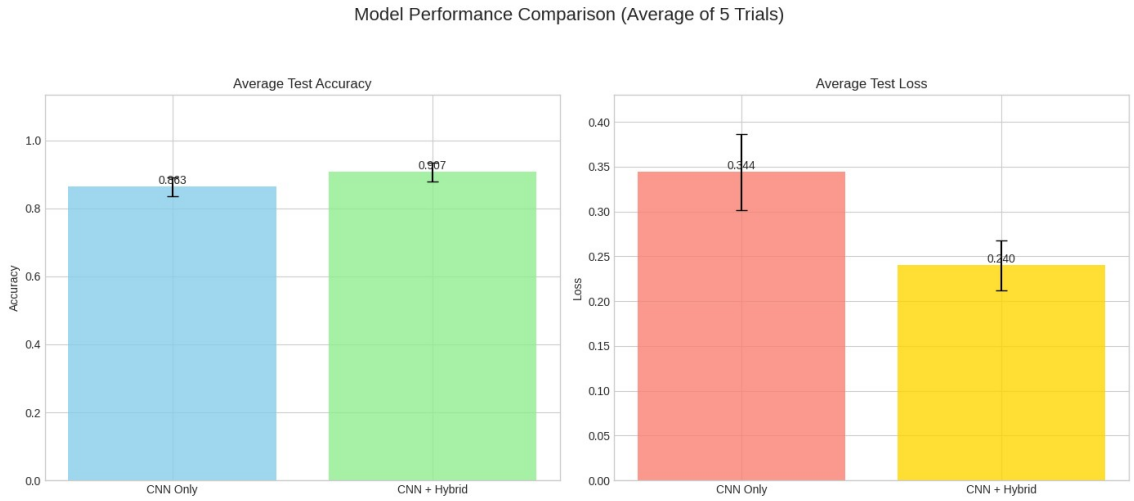



図 5.1：実験 1 のモデル性能比較（Accuracy / Loss）

実験 1 の結果、メタデータの追加により正解率が約 4.3%向上し、損失値も大幅に低下した。これはバックボーンを凍結したことで、メタデータ由来のノイズが画像特徴抽出能力を破壊することなく、画像情報を補完する有効な指標として機能したことを示している。

# 6. 実験 2：CNN vs ViT におけるマルチモーダル性能の分析

## 6.1 実験設定

実験 2 では、実験 1 と同じ合計 16 次元のメタデータベクトルを用いた。ただし、実験 1 が画像統計量に特化していたのに対し、実験 2 では画像統計量を 9 次元に絞り、新たに被写体属性（1 次元）と言語的特徴（3 次元）等を加えた計 16 次元（表 4.1 参照）とした。

本実験の重要な設定変更として、バックボーンの重み更新を許可し、全てのパラメータを学習可能とした。

統合したメタデータは、実験 1 の内容に加え、以下の高次な意味情報を含む全項目である。

- 顔の数（Face Count）：被写体属性
- 文字情報（Text/OCR）：言語的な意味内容

## 6.2 実験結果

各モデルにおける正解率および F1-score の比較を表 6.1 に示す。

表 6.1：実験 2 における各モデルの性能比較（バックボーン Unfrozen）

| Model Type | Modality | Mean Accuracy | F1-Score | 備考 |
|---|---|---|---|---|
| EfficientNet (CNN) | single | 87.33% | 83.25% | |
| EfficientNet (CNN) | hybrid | 86.00% | 76.80% | 性能劣化 |
| ViT | single | 91.67% | 89.95% | |
| ViT | hybrid | 92.00% | 93.33% | 最高精度 |

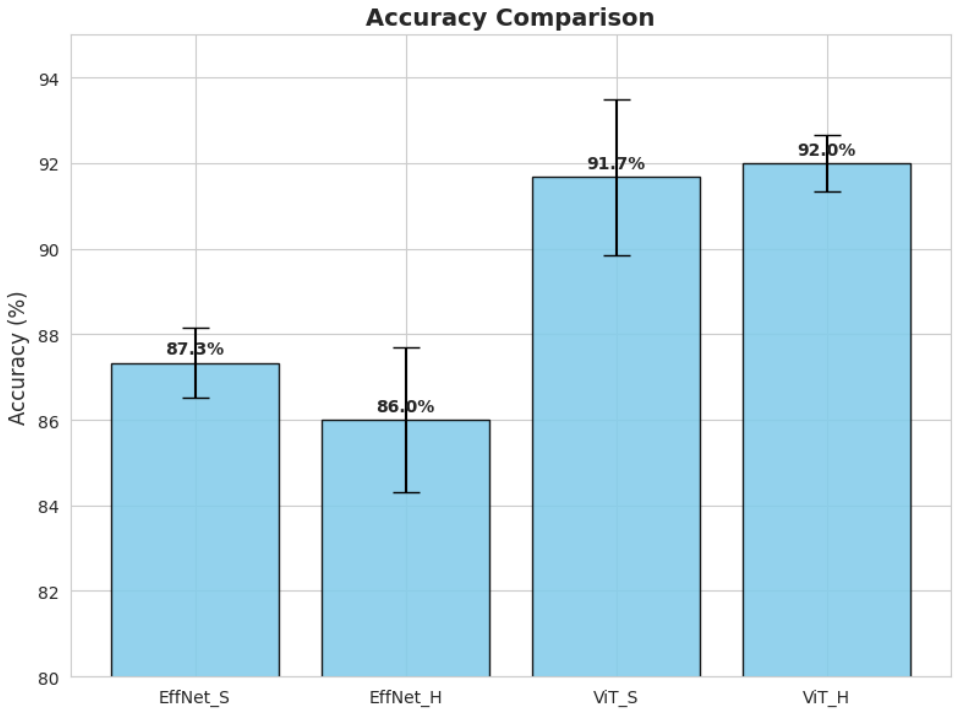



図 6.1：Accuracy の比較グラフ

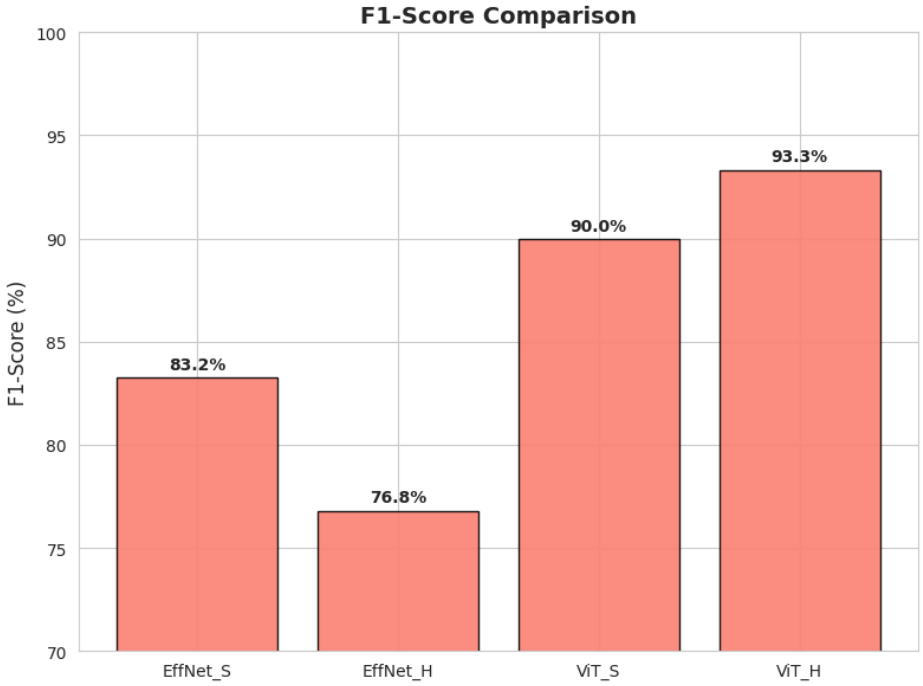
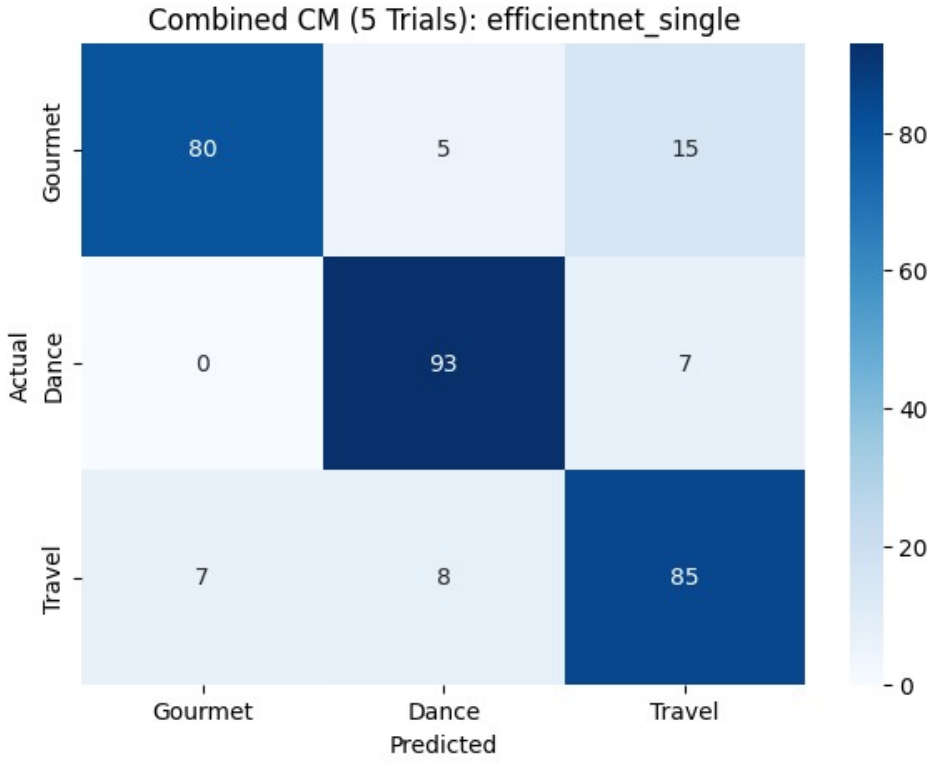
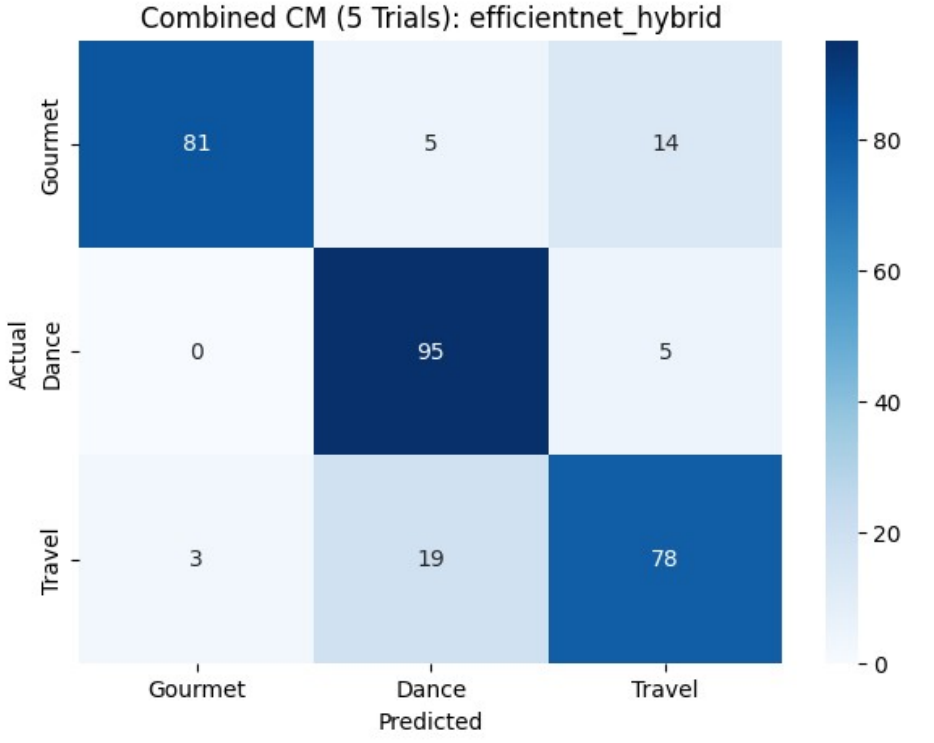



図 6.2：F1-Score の比較グラフ

## 6.3 混同行列

各モデルの分類傾向を可視化した混同行列を図 6.3 に示す。



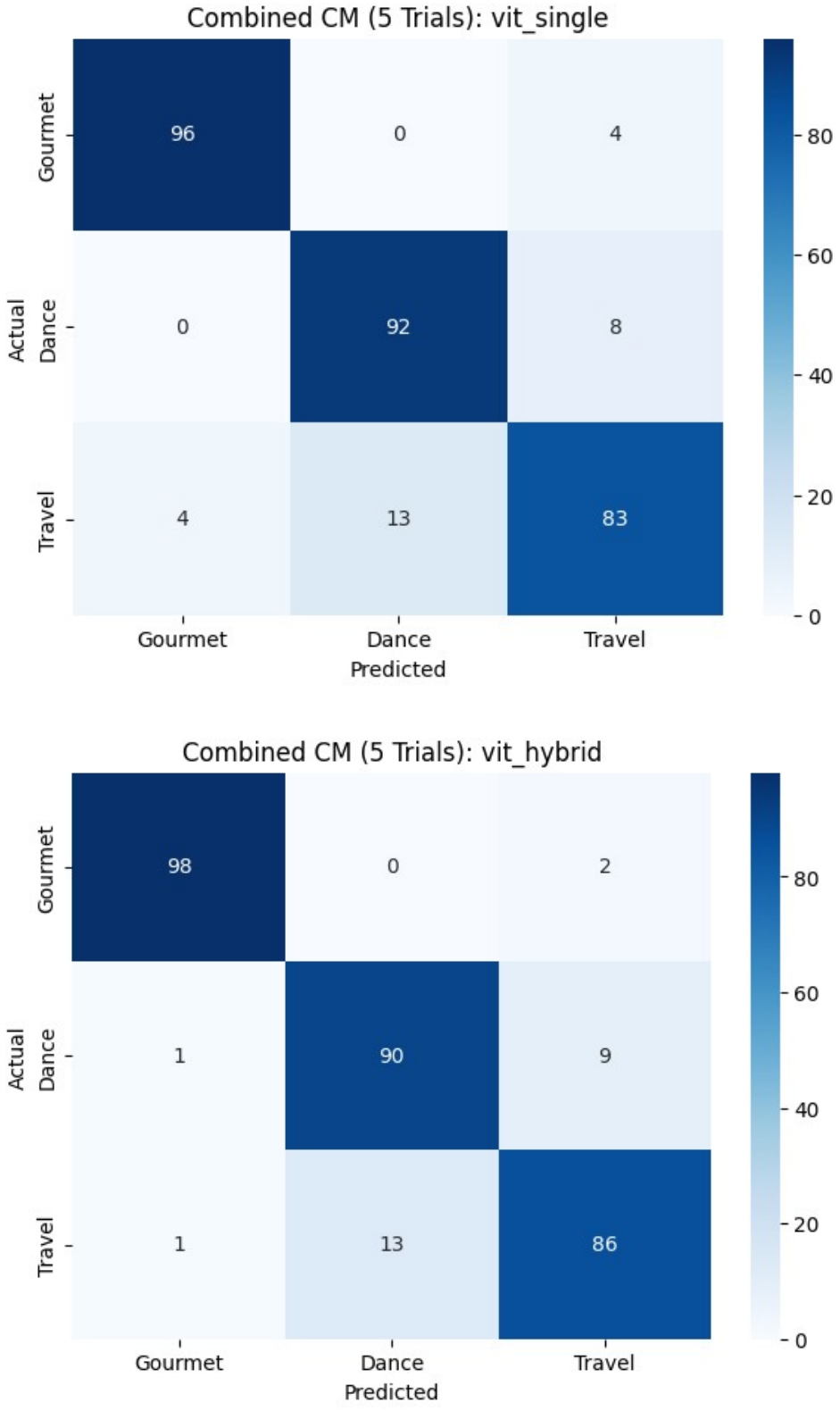



図 6.3：各モデルの混同行列（Confusion Matrix）

分析の結果、EfficientNet-Hybrid において顕著な精度の低下が確認された 。特に「Travel」クラスを「Dance」と誤認識するケースが激増しており、Hybrid 化が逆に識別境界を不安定化させている。対照的に ViT モデルでは、Hybrid 化によって「Gourmet」の正解数が 96 から 98 へ向上するなど、高次元データの追加が正の方向に作用し、F1-score 93.3%という最高精度を達成した。

実験 2 における CNN の精度低下は、メタデータの総次元数が原因ではなく、テキスト情報の 3 次元（text_len, has_text, keyword_count）等の意味的データが、全層の最適化を適用したバックボーンに対してノイズとして作用した結果である。

# 7. 結論と考察

本研究の実験結果に基づき、モデルアーキテクチャとメタデータの親和性、およびバックボーンの設定がマルチモーダル学習に与える影響について論じる。

## 7.1 CNN における特徴空間における干渉と勾配汚染

実験 1 において CNN の精度を向上させたメタデータが、実験 2 において逆に精度の低下を招いた要因は、バックボーンの学習可能性（Trainability）と、統合プロセスにおける次元的な性質に集約される。

### 7.1.1 バックボーンの全層ファインチューニングと勾配汚染

実験 1 ではバックボーンを凍結（Frozen）したことで、CNN は ImageNet で学習された高品質な特徴抽出器として安定して機能し、メタデータはその補助情報として作用した。しかし、実験 2 ではバックボーンを全層ファインチューニングしたため、結合部を通じてメタデータ由来の勾配が CNN 本体へ逆伝播する構造となった 。

第 4.3 節（表 4.2）の統計分析が示す通り、本研究で用いたインプレッションデータは標準偏差が平均値を大幅に上回っており（例：いいね数の $\sigma \approx 109$万 に対し $\mu \approx 19.7$万）、極めて変動率が高いノイズを含んだデータである。また、第 4.4.1 節の回帰分析において、「グルメ」カテゴリの保存数といいね数の決定係数が $R^2 = 0.090$ という極めて低い値を示した事実は、これらのメタデータがコンテンツのカテゴリに対して一貫した決定権（信号）を持たないことを裏付けている。

マルチモーダル学習における情報の質の不均衡（Modality Imbalance）に関する研究（Wang et al., 2020）によれば、画像情報のような強固な視覚的特徴と比較して、このような「弱い信号」しか持たないモダリティ由来の不安定な勾配は、学習プロセス全体を不安定化させることが指摘されている。さらに、Huang ら（2022）が定義した「モダリティ競合（Modality Competition）」の理論に基づけば、不適切なメタデータ情報の統合は、全層ファインチューニングされた CNN のフィルタを不適切に更新し、事前学習で得られた視覚特徴抽出能力を破壊する「勾配汚染」を引き起こす。実験 2 における CNN Hybrid の精度低下（87.33%から 86.00%への低下）は、まさにこの統計的ノイズが画像認識の最適化を阻害した結果であると断定できる。

### 7.1.2 特徴空間におけるモダリティ間の干渉と識別境界の不安定化

本研究の実装における情報の統合プロセスでは、Global Average Pooling（GAP）を用いて画像特徴マップの空間情報を 1 次元ベクトルへと圧縮した後に、メタデータ特徴を連結している 。

実験 2 における CNN の精度低下は、高次元特徴空間において「意味的性質の異なるベクトルを直接結合」したことに起因する干渉が主因であると考えられる。第 4.3 節の統計分析（表 4.2）が示す通り、社会的知覚データは標準偏差が平均値を大幅に上回る不安定な特性を有している 。このような統計的ノイズを含む「弱い信号」が、バックボーン全層ファインチューニング時の勾配更新プロセスにおいて画像由来の濃密な視覚特徴と同一のベクトル空間で演算されたことにより、識別境界の複雑化を招いたと推察される。

特に、Huang ら（2022）が提唱する「モダリティ競合（Modality Competition）」の観点から見れば、不安定なメタデータ由来の勾配が、全層の最適化を適用した CNN のフィルタを不適切に更新し、事前学習で得られた視覚特徴抽出能力を破壊する「勾配汚染」を引き起こした可能性が高い 。このように、CNN を用いた特徴レベルの後方融合においては、結合部におけるモダリティ間の質的乖離がモデル全体の最適化を阻害するリスクが顕在化したと言える 。

## 7.2 CNN と ViT のアーキテクチャの違い

実験 2 において、ViT のみがマルチモーダル化による安定した精度向上（91.67%から 92.00%）を示した要因は、CNN と ViT それぞれの内部表現の性質に起因する。

### 7.2.1 CNN の「特徴マップ」と局所性

EfficientNet に代表される CNN は、情報の局所性（Locality）と「テクスチャバイアス」に強く依存する 。空間構造を保持する特徴マップに対し、画像全体を要約する性質を持つメタデータを単純に連結したことで、結合特徴空間において分布の不連続性が生じ、識別境界の複雑化を招いたと考えられる 。

### 7.2.2 ViT における大域的結合と意味的親和性

対照的に、Vision Transformer は Vaswani ら (2017) が提案した Self-Attention 機構に基づき、画像内の全パッチが距離に関係なく直接的に相互作用を行う。

この大域受容野により、ViT は画像全体の「構成」や「意味的文脈」を捉える能力（Shape Bias）が高いことが知られている（Dosovitskiy et al., 2020）。
ViT の最終出力である [CLS] トークン（または GAP されたトークン）は、画像全体のパッチから重要な情報を動的に選択・集約したものであり、その表現はより抽象化された「意味の多様体」上に位置する。この「画像由来の意味情報」と「メタデータ由来の意味情報」は、共に高次な概念レベルで整合しており、高い意味的親和性（Semantic Affinity）を有していると考えられる。この親和性が、単純な Late Fusion であっても両者の情報を

滑らかに統合することを可能にし、ViT_H における最高精度の達成（Accuracy 92.0%, F1-score 93.3%）に寄与したものと考えられる。

## 7.3 モデルの大きさの違い

実験 2 において ViT が CNN を上回る精度を記録した要因の一つとして、前述した「モデル規模の圧倒的な差」が寄与している可能性を否定できない。約 8,600 万のパラメータを持つ ViT は、約 530 万の CNN と比較して情報の受容容量が大きく、メタデータに含まれる複雑な非線形関係やノイズを吸収・処理する能力において優位であったと考えられる。

また、本研究のデータセット（計 300 枚）という小規模な条件下での全層学習（Full Fine-tuning）においては 、本来であればパラメータ数の多いモデルほど過学習のリスクが高まる。しかし、ViT においては事前学習による強固な表現獲得と、大域的な Attention 機構 が、少規模データにおける不安定なメタデータの干渉を抑制するように働いたと推察される。

## 7.4 結論

本研究により、SNS 画像分類タスクにおけるマルチモーダル情報の統合において、ViT は CNN よりも優れた適性を示すことが明らかになった 。CNN は局所的なテクスチャ認識に優れる反面、抽象度の高い意味的メタデータの統合において情報の干渉や勾配汚染を招きやすく、精度の低下を招くリスクがある。一方、大域的な文脈理解を基本とする ViT は、全層ファインチューニングの設定下であっても異種データの統合による恩恵を享受しやすく、複雑なマルチモーダル情報の統合において極めて有効なアーキテクチャであることが示唆された 。

## 7.5 今後の課題

本研究の限界と今後の展望として、以下の 2 点が挙げられる。

1. モデル規模の統制: 今回は軽量な CNN（EfficientNet-B0）と標準的な ViT（ViT-B/16）を比較したが、今後は同程度のパラメータ数を持つモデル（例：EfficientNet-B7 や ViT-Tiny 等）を用いることで、純粋なアーキテクチャ由来の特性をより厳密に検証する必要がある。
2. データセットの拡充: 300 枚という限定的なデータセットでは 、バックボーン全層の最適化を適用した時の勾配汚染の影響が過大に現れる可能性がある。より大規模

なデータを用いた検証を行うことで、マルチモーダル統合における CNN の課題が「データ量による解決が可能か」あるいは「構造的な限界か」を切り分けることができるだろう。

# 8. 参考文献


1. 山本晋太郎, 徳井直生, ほか. 『Vision Transformer 入門 (Computer Vision Library)』, 技術評論社, 2022 年。
2. 東京大学工学系研究科. 「グローバル消費インテリジェンス寄附講座 (GCI) 2024 Winter 講義資料」, 2024 年。
3. 東京大学工学系研究科. 「Deep Learning 基礎講座 2025 Spring 講義資料」, 2025 年。
4. 斎藤康毅. 『ゼロから作る Deep Learning —Python で学ぶディープラーニングの理論と実装』, オライリー・ジャパン, 2016 年。
5. 一般社団法人日本ディープラーニング協会. 『深層学習教科書 ディープラーニング G 検定（ジェネラリスト）公式テキスト 第 2 版』, 翔泳社, 2021 年。
6. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11), 2278-2324。
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is all you need." Advances in Neural Information Processing Systems, 30。
8. Van Rijsbergen, C. J. (1979). Information Retrieval (2nd ed.). Butterworth-Heinemann。
9. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." Advances in Neural Information Processing Systems, 25。
10. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." arXiv preprint arXiv:1811.12231。
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929。
12. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). "Learning deep features for discriminative localization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2921-2929。
13. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press。
14. Tan, M., & Le, Q. (2019). "EfficientNet: Rethinking model scaling for convolutional neural networks." International Conference on Machine Learning, 6105-6114。

15. Eagleman, D. M. (2001). "Visual illusions and neurobiology." Nature Reviews Neuroscience, 2(12), 920-926。
16. Huang, Y., Lin, J., Zhou, C., Hong, Y., & Pan, S. J. (2022). "Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning?" arXiv preprint arXiv:2203.03305。
17. Wang, W., Tran, D., & Feiszli, M. (2020). "What Makes Training Multi-modal Classification Networks Hard?" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12695-12705。
18.

# 9. 謝辞

本論文の完成にあたり、多くの方々から多大なるご支援とご指導を賜りました。

指導教員である Ruggero Micheletto 教授には、研究の進め方から論文の執筆、さらには図表の科学的な作成方法に至るまで、多岐にわたるご指導をいただきました。特に、研究室のサーバを利用したプログラミングの指導を通じて、技術的な研鑽を積む機会をいただいたことに深く感謝申し上げます。

研究室の先輩である河野さんおよび小泉さんには、3 年次の研究テーマ模索期において、錯視の自作に関する貴重な助言をいただきました。研究室の運営や就職活動においても、常に穏やかかつ親身に寄り添っていただいたおかげで、困難な時期を乗り越えることができました。

また、及川さんには毎週のゼミでのディスカッションを通じて、新たな視点や研究の方向性について多くの補助をいただきました。同期の武田隼人さんと市川凜さんには、研究への真摯な姿勢に常に刺激をもらい、互いに切磋琢磨しながら研究に励むことができました。心より感謝いたします。

最後に、日頃の生活を支えてくれた家族、そして大学生活を共にした友人たちに深い感謝の意を表します。

木村 裕健

# 10. 付録

## 10.1 第 5 章 実験 1：CNN における低次元メタデータの統合評価用プログラム

実験 1 で使用した、EfficientNet-B0 のバックボーンを固定し、画像統計量と「いいね数」を統合するモデル（TensorFlow/Keras 版）の実装である

```
import os

import tensorflow as tf

from tensorflow.keras import layers, models

from tensorflow.ker

as.preprocessing import image_dataset_from_directory

import numpy as np

import pandas as pd

from sklearn.preprocessing import StandardScaler

# ========================================

# 1. 実験環境および共通設定

# ========================================

NUM_TRIALS = 5

EPOCHS = 15

BATCH_SIZE = 32

IMG_HEIGHT = 224

IMG_WIDTH = 224
```

```
# ========================================

# 2. モデルアーキテクチャ定義

# ========================================

def build_cnn_single_model(num_classes):

"""画像情報のみを用いる単体CNNモデルの構築"""

# データ拡張層

data_augmentation = models.Sequential([

layers.RandomFlip("horizontal"),

layers.RandomRotation(0.1),

layers.RandomZoom(0.2)

], name='data_augmentation')

# 特徴抽出部（EfficientNetB0）

base_model = tf.keras.applications.EfficientNetB0(

include_top=False, weights='imagenet', input_shape=(IMG_HEIGHT, IMG_WIDTH, 3)

)

base_model.trainable = False # バックボーンの凍結（実験１の設定）

inputs = layers.Input(shape=(IMG_HEIGHT, IMG_WIDTH, 3))

x = data_augmentation(inputs)

x = tf.keras.applications.efficientnet.preprocess_input(x)

x = base_model(x, training=False)
```

```
x = layers.GlobalAveragePooling2D()(x)

# 分類層

x = layers.Dense(128, activation='relu')(x)

x = layers.Dropout(0.3)(x)

outputs = layers.Dense(num_classes, activation='softmax')(x)

return models.Model(inputs, outputs)

def build_hybrid_model(num_classes, meta_dim):

"""画像とメタデータを特徴レベルで統合するHybridモデルの構築"""

# 1. 画像処理ブランチ

image_input = layers.Input(shape=(IMG_HEIGHT, IMG_WIDTH, 3), name='image_input')

base_model_h = tf.keras.applications.EfficientNetB0(

include_top=False, weights='imagenet', input_tensor=image_input

)

base_model_h.trainable = False # バックボーンの凍結

image_features = layers.GlobalAveragePooling2D()(base_model_h.output)

# 2. メタデータ処理ブランチ（MLP）

meta_input = layers.Input(shape=(meta_dim,), name='meta_input')

meta_features = layers.Dense(64, activation='relu')(meta_input)

meta_features = layers.Dense(32, activation='relu')(meta_features)

# 3. 特徴レベルの統合（Late Fusion）
```

```
combined = layers.concatenate([image_features, meta_features])

# 4. 共通分類層

x = layers.Dropout(0.5)(combined)

x = layers.Dense(128, activation='relu')(x)

output = layers.Dense(num_classes, activation='softmax')(x)

return models.Model(inputs=[image_input, meta_input], outputs=output)

# ==========================================

# 3. 学習プロセス

# ==========================================

def train_model(model, train_ds, test_ds, is_hybrid=False):

"""モデルのコンパイルおよび学習の実行"""

loss_fn = 'sparse_categorical_crossentropy' if is_hybrid else 'categorical_crossentropy'

model.compile(

optimizer='adam',

loss=loss_fn,

metrics=['accuracy']

)

history = model.fit(

train_ds,

validation_data=test_ds,
```

```
epochs=EPOCHS,

verbose=1

)

return history
```

## 10.2 第 6 章 実験 2：CNN vs ViT におけるマルチモーダル性能の分析用プログラム

実験 2 で使用した、CNN（EfficientNet）と ViT（Vision Transformer）を切り替えて比較可能な、全層ファインチューニング用モデル（PyTorch 版）の実装である 。

```
import os

import torch

import torch.nn as nn

import torch.optim as optim

from torch.utils.data import Dataset, DataLoader

from torchvision import models, transforms

from PIL import Image



# ==========================================

# 1. 実験設定（Hyperparameters）

# ==========================================

NUM_TRIALS = 5

EPOCHS = 15
```

```
BATCH_SIZE = 32

DEVICE = torch.device("cuda" if torch.cuda.is_available() else "cpu")



# ========================================

# 2. モデルアーキテクチャ（Late Fusion）

# ========================================

class MultiModalClassifier(nn.Module):

    """CNN/ViTとメタデータを特徴レベルで統合する分類器"""

    def __init__(self, model_type='efficientnet', num_meta=16, is_hybrid=True):

        super().__init__()

        self.is_hybrid = is_hybrid



        # 画像バックボーンの選択

        if model_type == 'efficientnet':

            self.backbone =
models.efficientnet_b0(weights=models.EfficientNet_B0_Weights.DEFAULT)

            self.img_dim = self.backbone.classifier[1].in_features

            self.backbone.classifier = nn.Identity()

            # 実験2では self.backbone.parameters() を学習対象とする（Unfrozen）

        else: # ViT-B/16

            self.backbone = models.vit_b_16(weights=models.ViT_B_16_Weights.DEFAULT)
```

```
    self.img_dim = self.backbone.heads.head.in_features
    self.backbone.heads = nn.Identity()

# メタデータ処理層（Embedding）
if is_hybrid:
    self.meta_layer = nn.Sequential(
        nn.Linear(num_meta, 64),
        nn.ReLU(),
        nn.Dropout(0.2)
    )
    input_dim = self.img_dim + 64
else:
    input_dim = self.img_dim

# 最終分類器（Classifier）
self.classifier = nn.Sequential(
    nn.Linear(input_dim, 256),
    nn.ReLU(),
    nn.Dropout(0.3),
    nn.Linear(256, 3)
```

```
        )

    def forward(self, img, meta):
        # 画像特徴抽出
        x_img = self.backbone(img)

        # 特徴レベルでの統合
        if self.is_hybrid:
            x_meta = self.meta_layer(meta)
            x = torch.cat((x_img, x_meta), dim=1)
        else:
            x = x_img

        return self.classifier(x)

# ==========================================
# 3. 学習および評価ループ
# ==========================================
def run_trial(model, train_loader, val_loader):
    """各試行における学習（全層ファインチューニング）および検証プロセスの実行"""
```

```
criterion = nn.CrossEntropyLoss()

# バックボーンを含む全パラメータを最適化対象とする

optimizer = optim.Adam(model.parameters(), lr=1e-4)

best_acc = 0



for epoch in range(EPOCHS):

    model.train()

    for imgs, metas, labels in train_loader:

        imgs, metas, labels = imgs.to(DEVICE), metas.to(DEVICE), labels.to(DEVICE)

        optimizer.zero_grad()

        outputs = model(imgs, metas)

        loss = criterion(outputs, labels)

        loss.backward()

        optimizer.step()



    # 検証（Evaluation）プロセスは省略（ベスト精度の保存）

return best_acc
```

11 QA 対策

1. なぜ F1score を使用するのか？不均衡データではなく、重要性は低いことは承知だが、モデルを多方面から評価することは意味がある。偽陽性と偽陰性がわかることにより、間違いの質が分かる。
2. LeCun (1998) 等の先行研究では誤り率(1-Accuracy)による評価が一般的であったが、本稿では現代の画像認識研究の慣例に従い、正解率を用いて表記する。
3. 


(ア) Vision Transformer 入門 (Computer Vision Library)

(イ) 東京大学大学院工学系研究科技術経営戦略学 グローバル消費インテリジェンス寄附講座 2024 winter

(ウ) 東京大学大学院工学系研究科技術経営戦略学 Deep Learning 基礎講座 2025 Spring

(エ) ゼロから作る Deep Learning - O'Reilly Japan

(オ) 深層学習教科書 ディープラーニング G 検定(ジェネラリスト)公式テキスト 第 2 版

(カ) LeCun, Y., et al. (1998): "Gradient-based learning applied to document recognition." CNN 畳み込み演算

(キ) Vaswani, A., et al. (2017): "Attention is all you need." Attention 機構

(ク) Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Butterworth-Heinemann. F1-score の引用

(ケ) Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." Accuracy の引用

(コ) Geirhos, R., et al. (2019). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." CNN が「形状」ではなく「テクスチャ」に依存するという「Texture Bias」を論証した

(サ) Dosovitskiy, A., et al. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale." Vision Transformer（ViT）を最初に提案した論文です。ViT が CNN と異なり、大域的な情報を初期層から扱えることを述べる

(シ) Zhou, B., et al. (2016). "Learning deep features for discriminative localization." Global Average Pooling (GAP) が空間情報をどのように圧縮し、特徴マップを生成するか

(ス) Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Broadcasting

(セ) Tan, M., & Le, Q. (2019). "EfficientNet: Rethinking model scaling for convolutional neural networks." EfficientNet の原著と構造

(ソ) Eagleman, D. M. (2001). "Visual illusions and neurobiology. " 錯視と脳構造、立命館も引用

(タ) Huang et al. (2022): Huang, Y., et al. "Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning?" マルチモーダル化して精度が下がること

(チ) Wang et al. (2020): Wang, W., et al. "What Makes Training Multi-modal Classification Networks Hard?" モダリティごとの精度の差(インバランス)がモデルを弱くする

4.

5.