CH-002

雲画像とキャプション生成を統合した雲形分類モデルの検討

Investigation of a Cloud Type Classification Model Integrating Cloud Images and Caption Generation

西野 大河 † 矢野 耕太郎 ‡ 遠藤 聡志 § 武井 弘樹 ¶
Taiga Nishino† Kotaro Yano‡ Satoshi Endo§ Hiroki Takei¶

1 はじめに

雲は気象変化を把握する上で重要な指標であり、特に積乱雲の発生は豪雨など急激な天候変化の前兆となる。雲形を正確に分類し、降水の有無やその変化を予測できる。日中のドライブレコーダー画像を用いた地上カメラ観測は、局地的な雲の形状や動きをリアルタイムで捉えることができ、数分後の降水予測により、特に沖縄に多い突発的なにわか雨（カタブイ）を効果的に把握できる。衛星観測は広域の状況把握には優れる一方、局所的な変化には対応しにくく、地上カメラによる観測はこうした局地的現象の検出に適している。沖縄は高温多湿な気候に加え、地形や海洋の影響により、雨をもたらす雲の発生に伴う急激な天候変化が生じやすい地域である。故に、雲の観測と分類は本地域において特に重要である。

末光らの研究 [1] では、降水をもたらす積乱雲と乱層雲、その他の雲をまとめて 1 クラスとする 3 クラス分類を行い、Jinglin ら [2] が構築した CCSN データセットを使用した。画像中の雲以外の領域（地上や建物など）が分類に与える悪影響を低減するため、Semantic Segmentation モデル DeepLabv3+ を用いて雲と空の領域のみを抽出し、そのマスク画像を利用して CNN の一種である EfficientNetB0 で学習を行う手法を提案した。この手法により、5 分割交差検証における test に対する平均精度が 68.61% まで向上したが、マスク精度の不安定さが依然として課題であった。以上のことから、分類精度をさらに向上させるには、マスクに依存しない新たなアプローチの導入が必要であると考えられる。

近年では、画像キャプション生成が画像分類タスクの性能向上に寄与することが報告されている。たとえば、BLIP2 を用いたキャプション生成文をデータセットに付加することで、ImageNet の分類精度が 4.4% 向上しており [3]、Vision-Language Model (VLM) の有効性が示唆される。また、マルチモーダル情報を統合的に処理するモデルにおいても高い分類性能が報告されており、Gao et al. [4] では、画像とテキストの階層的特徴表現を単純な特徴ベクトルの結合により統合する深層マルチモーダル融合手法により、分類精度と検索タスクにおいて単一モーダルを上回る性能を実現している。

以上の研究成果を踏まえ、本研究では、雲画像に対する分類タスクにキャプション生成を組み合わせたマルチモーダル手法を導入し、雲形分類精度のさらなる向上を図る。具体的には、画像とその説明文（キャプション）を同時に扱うモデルを構築し、従来手法を超える分類性能を実現すること、およびキャプション生成を取り入れた分類手法の有効性を検証することを目的とする。

† 琉球大学大学院理工学研究科知能情報プログラム,Graduate School of Engineering and Science, University of the Ryukyus
‡ 琉球大学大学院理工学研究科知能情報プログラム,Graduate School of Engineering and Science, University of the Ryukyus
§ 琉球大学工学部工学科知能情報コース, Computer Science and Intelligent Systems, University of the Ryukyus
¶ 株式会社ウェザーニューズ, Weathernews Inc.

2 提案手法

図 1 に本提案手法の全体構成を示す。本研究では、画像とテキストを統合的に用いたマルチモーダルな手法によって、雲形分類精度の向上を図る。従来の画像分類は視覚的特徴のみに基づいており、雲形のように視覚的差異が微細な対象では、分類性能に限界があると考えた。これに対し、本研究では、雲画像から生成されるキャプション生成文を意味的情報として活用し、画像情報との統合によって精度向上を目指す。

まず、雲画像に対してキャプション生成モデルを適用し、雲の視覚的な特徴を詳細に記述したキャプション生成文を得る。このキャプション生成文のみを入力とし、事前学習済みの BERT を雲に関する説明文でファインチューニングしたモデルを用いて意味的特徴を抽出する。そして、BERT によって得られた特徴ベクトルを用いて雲形分類を行う。

さらに、テキスト特徴と画像特徴を統合したマルチモーダルな分類を実施する。具体的には、上記の BERT によるテキスト特徴に加え、CNN により抽出された画像特徴を用いる。CNN は雲画像を対象に本研究でファインチューニングを行っており、画像情報に対する識別能力を高めている。これらテキスト特徴と画像特徴を結合し、多層パーセプトロン（MLP）に入力することで、両モダリティを統合した分類を実現する。

3 データセットの構築

3.1 使用するデータセット

使用するデータセットは先行研究で使用された CCSN データセットから自然発生する雲ではない Ct(飛行機雲) クラスを除いた、Cb(積乱雲) クラス 242 枚、Ns(乱層雲) クラス 274 枚、およびその他 8 種類のラベルを 1 クラスとした 226 枚、計 742 枚の画像を使用する。これを train:validation:test = 6:2:2 になるように分割する。

3.2 キャプション生成とプロンプト設計

本研究では、詳細な画像説明文の生成能力と多様な視覚的特徴の捉え方に優れる大規模言語モデルの特性に着目し、CCSN データセットの雲画像に対し GPT-4.1-mini を用いてキャプションを生成した（プロンプトおよび出力は英語）。これらのキャプション生成文は、雲形分類における有効な入力特徴となるよう、次の段階的なプロセスに基づきプロンプトを設計した。

まず、プロンプト設計の初期段階として、代表的な 10 種類

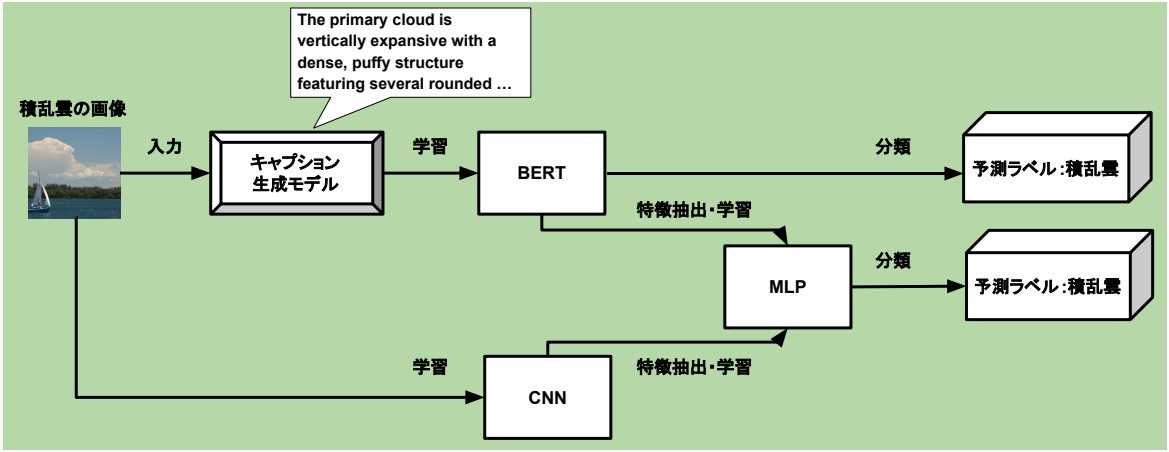

図 1: 提案手法の全体構成

の雲の典型的な視覚的特徴を GPT-4o に箇条書き形式で説明させた。これにより、雲形分類に必要とされる構造的・光学的な観察項目（例：頂部の明瞭さ、底部の高さ、層状性、照明の方向など）を明確化し、プロンプト設計に反映させた。その上で、キャプション文の品質と一貫性を担保するため、表 1 に示すプロンプト制約を設けた。これらの制約は、キャプションが分類に有効な視覚的・構造的情報を安定的に含むように意図されたものである。例えば、頂部の明瞭さや底部のぼやけ具合、光の向きといった特徴は、雲種の推定にとって重要な判断材料となりうる。また、文構造を統一することで、BERT の分類モデルにおいて文の位置情報を利用した学習の一貫性が高まり、精度向上が期待できる。

このように設計されたプロンプトに基づくキャプション生成は、「キャプションの評価 → プロンプトの修正 → 再生成 → 再評価」という反復的な工程を通じて最適化された。最終的には、分類精度の向上に寄与する情報構造を含んだ、定型的かつ分類指向のキャプション文が得られたと判断する。

3.3 キャプション生成の結果

キャプション生成文の出力例を積乱雲の画像とそのキャプション生成文の例として図 2 に示す。この例では、積乱雲の一般的な特徴がキャプション生成文に再現されている。主雲が縦に広がり、密度が高く、膨らんだ構造を持つことが記述されており、照明効果や影の発生、雲の大きさや空間的位置、頂部と底部の特徴が客観的に表現されている。特に、頂部が明瞭で底部がやや不明瞭であること、背景の空が雲によって覆われていることなど、分類に有用な情報が過不足なく含まれている点が注目される。なお、GPT-4.1-mini によるキャプションの平均語数は、1 画像あたり 75.96 語であった。

"The primary cloud is vertically expansive with a dense, puffy structure featuring several rounded protrusions. It is predominantly white with subtle shading on some of its bulges, indicating gentle lighting variations and slight shadowing. The cloud appears very large and extends high above the horizon, occupying much vertical space in the sky. Its top is well-defined and crisp, while the base is less distinct but appears elevated. The cloud blocks much of the sky behind it, suggesting a thick and opaque composition."

図 2: 積乱雲の画像とそのキャプション生成文の例

表 1: プロンプトにおける記述制約（日本語訳）

1. 主対象の明確化：画像内で最大の雲のみに焦点を当て、他の雲や地上の要素、時間帯に関する記述は除外する。
2. 雲種名の回避：積乱雲などの具体的な分類名は使用しない。
3. 5 つの記述項目の順守（a〜e）：
   （a）形状と構造
      - 内容：雲の全体的な形状および構造的特徴（例：層状、膨らみ、平坦など）を記述。
      - 例文：（略）
   （b）色と光の効果
      - 内容：雲の色や照明効果（影、逆光など）について記述。
      - 例文：（略）
   （c）大きさと空における位置
      - 内容：視覚的な大きさおよび空の中での相対的位置を記述。
      - 例文：（略）
   （d）頂部と底部の特徴
      - 内容：頂部や底部の明瞭さ、輪郭、位置の高さなどを記述。
      - 例文：（略）
   （e）視覚的に観察可能な天気状況
      - 内容：雲の見た目から明確に推測できる周囲の天気を記述（主観的な印象は避ける）。
      - 例文：（略）
4. 分類に有用な特徴も可能な範囲で記述：例として、頂部の明瞭さ、底部の高さ、照明の向きなどを含める。
5. 事実に基づく客観的な表現：主観的や曖昧な語句を避け、「主雲」に帰属する特徴を具体的に記述。
6. 文体と分量の制約：文数は 4〜5 文、語数は最大 80 語。簡潔で冗長を避け、記述項目 a〜e の順序で構成される。
7. 空や天気の描写の制限：雲の構造や光学的特徴と直接関係する場合のみ、背景や天気について記述可。

4 実験設定

本研究では、雲形分類におけるキャプション生成文の有効性を検証するため、画像、キャプション生成文、および画像とキャプション生成文を組み合わせた入力データを用いた分類性能比較を実施する具体的には、それぞれ CNN、BERT、そして MLP を基盤としたモデルを構築し、各モデルの分類性能を詳細に比較・評価する。

4.1 モデル構築と学習の最適化

本研究では、それぞれの分類タスクに対し、最適なモデルアーキテクチャと学習設定を追求した。すべての分類モデルの最適化手法には、汎化性能と安定性のバランスに優れる AdamW を共通して採用した。また、各分類モデルはそれぞれのタスクに合わせた学習設定に基づき、最適なハイパーパラメータ探索を行った。最適化によって得られた最良のパラメータを用いて、画像分類、テキスト分類、マルチモーダル分類の各モデルの分類器を構築した。全ての実験は 5 分割交差検証により実施し、過学習の抑制のため Train Loss を監視対象とする Early Stopping を導入した。具体的なモデル構築とハイパーパラメータ最適化の詳細を以下に述べる。

4.1.1 画像分類モデル

画像分類では、CCSN データセットの雲画像で事前学習済みの EfficientNetV2-B0 を用いたファインチューニングを行った。訓練データには、過学習抑制と汎化性能向上のため、以下のデータ拡張を適用した。具体的には、50% の確率での水平フリップ、60% の確率での± 30 度回転、および± 50% 範囲での平行移動をランダムに実施した。バッチサイズは 8 と 16 でハイパーパラメータ最適化を行った。この CNN モデルは、先行研究で最も高い分類精度を示したグリッドサーチによるハイパーパラメータ設定を引き継いで学習を実行した。

4.1.2 テキスト分類モデル

テキスト分類では、CCSN の各雲画像に対応するキャプション生成文の分類を行った。前処理として、句読点除去、空白の統一、小文字化といった標準的なテキスト正規化を適用した。使用モデルは、事前学習済みの DistilBERT をファインチューニングを行った。バッチサイズは 16, 32, 64 の範囲で最適化を行った。

4.1.3 マルチモーダル分類モデル

マルチモーダル分類には、事前に雲画像および雲のキャプション生成文でそれぞれファインチューニングした EfficientNetV2-B0 と DistilBERT の重みを freeze し、特徴抽出器として使用した。これらを統合する Multi-layer Fusion Network を組み合わせたアーキテクチャを採用し、バッチサイズは 8 と 16 で最適化を行った。

特徴統合には単純な特徴ベクトルの結合手法を用いた。これは新たに導入したキャプション生成手法の有効性を基本的な構成で検証することを優先したためである。具体的には、抽出した画像特徴（1280 次元）とテキスト特徴（768 次元）に各々 LayerNorm を適用後、Linear 層で 256 次元に統一し、水平連結により 512 次元の融合特徴を生成する。その後、2 層 MLP で段階的に次元削減（512→256→128→3）を行い、三値分類を実施する。なお、LayerNorm は各サンプルに対して独立に正規化を行うため、画像情報とテキスト情報といった異なるモダリティ間でも一貫性のある正規化が可能である。

5 結果と考察

5.1 全体的な性能比較

本研究の 5 分割交差検証の結果を図 3 と図 4 に示す。表 2 に示すように、各手法の平均精度において、テキスト分類手法が 68.96% で最も高い精度を示した。一方で、単一 Fold における最高精度は僅差であり、テキスト分類手法が 70.90% でわずかに優位であった。これらの結果から、単なる最高値ではなく、各手法の安定性が平均精度に大きく影響していることが示唆される。

表 2: 各分類手法の精度比較（5 分割交差検証）

分類手法	平均精度 (%)	最高精度 (%)
画像分類手法	67.41	70.89
テキスト分類手法	**68.96**	**70.90**
マルチモーダル分類手法	67.46	70.15

5.2 各分類手法の詳細分析

5.2.1 画像分類手法の結果分析

画像分類手法の混同行列（図 3 左）を見ると、Cb（積乱雲）クラスは 82% と比較的高い正答率を示す一方で、Ns（乱層雲）クラスは 48% と低い正答率に留まっている。これは、Ns が他の層状雲（As, Cs, St など）と視覚的に類似しており、地上からの観測では層状雲間の区別が困難であることを示している。これらの結果は、視覚的に類似した層状雲の識別には、画像情報だけでは限界があることを示している。

5.2.2 テキスト分類手法の結果分析

テキスト分類手法の混同行列（図 3 中央）に注目する。テキスト分類は最も高い平均精度を示し、Ns（乱層雲）の正答率が 72% と、画像分類の 48% から大幅に改善されている点が特筆される。その理由として、DistilBERT が自然言語処理に特化したモデルであり、雲画像に付与されたキャプション文から、画像では判別しにくい「筋状」「広がりがある」といった構造的・文脈的特徴の抽出において高い性能を発揮したためと考えられる。また、画像分類よりもテキスト分類の精度が優位であることから、本研究で設計したプロンプトと GPT-4.1-mini によるキャプション生成が、雲形分類に有用な高品質な情報を適切に捉えられていたことが示唆される。すなわち、視覚的に区別が難しい雲形に対しても、言語情報が補完的な役割を果たしうる可能性を示すものである。一方で、「Other」クラスの Cb への誤分類が 26% と画像分類の 11% と比較して増加しており、テキストベースでは、「Other」クラスの判別が難しい傾向が見られる。これは、BERT が文脈を捉えることに長けている一方、「Other」クラスが 8 種類の異なる雲形をまとめた多様な集合体であるため、統一されたテキスト的特徴を抽出・学習することが困難であったためと考えられる。各雲形が持つ固有のテキスト特徴が互いに打ち消し合うことや、特定の表現が他クラスとの混同を招くことが示唆される。また、図 4 の箱ひげ図からも、テキスト分類は各 Fold 間でのばらつきが小さく、全体として非常に安定した分類性能を発揮していることが確認できる。

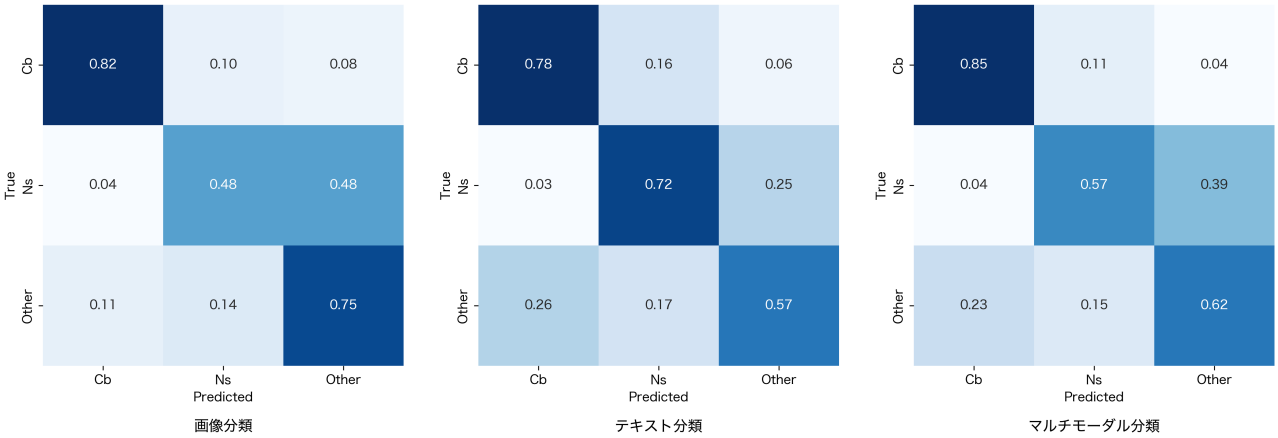

図 3: 各分類手法における混同行列（5 分割交差検証平均）

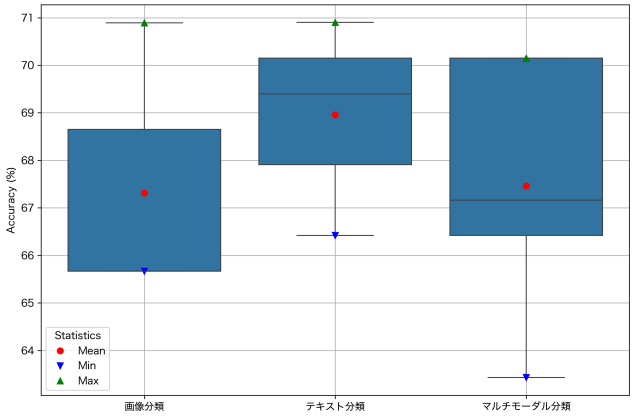

図 4: 各分類手法における精度分布（5 分割交差検証全結果）

5.2.3 マルチモーダル分類手法の結果分析

マルチモーダル分類手法の混同行列（図 3 右）を見ると、その性能は画像分類とテキスト分類のそれぞれの特徴を反映しており、両者の精度を足して 2 で割ったような結果となっている。Ns の正答率は 57% と画像分類の 48% から改善されているものの、テキスト分類の 72% には及ばない。これは、MLP が CNN（画像特徴）および BERT（テキスト特徴）の両方から抽出した特徴量を統合して学習しており、異なるモダリティの補完的な情報を活用できたためと考えられる。しかし、それぞれのモダリティで十分にファインチューニングされたにもかかわらず、単純な特徴ベクトルの結合だけでは、モダリティ間の意味的関連性を十分に活かしきれなかった可能性がある。このことから、マルチモーダル分類は、個々のモダリティの性能を単純に上回るものではなく、期待されたほどの性能向上が得られなかったと言える。これは、提案手法における特徴結合の方式（MLP による単純な特徴ベクトルの結合）が、複雑な雲形分類においてモダリティ間の情報を最適に統合するには不十分であったためと考えられる。図 4 の箱ひげ図から、マルチモーダル分類は Fold 間での精度のばらつきが大きく、特に最低値（Min）が他のモデルよりも低いことが確認できる。このばらつきは融合された特徴表現の不安定さや、過学習の影響によるものと考えられる。

6 まとめと今後の展望

本研究では、GPT-4.1-mini で生成した雲画像キャプションを雲形分類に適用し、画像、テキスト、マルチモーダルの分類性能を比較した。その結果、テキスト分類が最も高い分類精度を達成し、平均精度は 68.96% を記録した。これは先行研究のマスク手法による雲形分類の平均精度 68.61% を 0.35% 上回り、従来の画像のみの分類手法よりも高い精度であることが示した。一方、マルチモーダル分類については期待した性能向上に至らず、平均精度は 67.46% に留まった。これは、単純な特徴ベクトルの結合による融合手法では、各モダリティの情報を最適に統合できないことを示唆している。

今後の展望として、マルチモーダル分類の性能向上を目指し、より洗練された特徴融合手法の導入を検討する。具体的には、cross-attention 機構を用いたモダリティ間の相互作用を考慮した融合方法や、CLIP[5] のような画像とテキストの意味的な関連性を事前学習で獲得したモデルの活用が有効と考えられる。これらの手法により、各モダリティの補完的な情報をより効果的に統合し、雲形分類精度の向上が期待される。また、雲形分類における言語情報の有効性が検証されれば、実用的な気象観測システムへの応用可能性も期待される。

謝辞

本研究は JSPS 科研費 23K11234 の助成を受けたものである.

参考文献

[1] K. Suemitsu et al., ”Classification of Rainfall Intensity and Cloud Type from Dash Cam Images Using Feature Removal by Masking,” Climate, vol. 12, no. 5, p. 70, 2024.
[2] J. Zhang et al., ”CloudNet: Ground-based cloud classification with deep convolutional neural network,” Geophysical Research Letters, vol. 45, no. 16, pp. 8665–8672, 2018.
[3] T. Nguyen et al., ”Improving multimodal datasets with image captioning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 22047–22069, 2023.
[4] J. Gao et al., ”A survey on deep learning for multimodal data fusion,” *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.
[5] A. Radford et al., ”Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*, pp. 8748–8763, July 2021, PMLR.