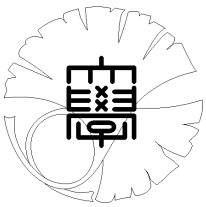
数理科学実践研究レター　2024–4 October 23, 2024

大規模マルチモーダル **AI** の数理と災害分野への応用

by

市川 佑馬、矢野 良輔



UNIVERSITY OF TOKYO

GRADUATE SCHOOL OF MATHEMATICAL SCIENCES

KOMABA, TOKYO, JAPAN

大規模マルチモーダル AI の数理と災害分野への応用

市川佑馬[1]（東京大学大学院総合文化研究科, 富士通人工知能研究所）
Yuma Ichikawa (Graduate School of Arts and Sciences, University of Tokyo, Fujitsu Limited)

矢野良輔[2] (東京海上ディーアール (株))
Ryosuke Yano (Tokio Marine dR Co., Ltd.)

概 要

本研究では, オープンソースの災害関連画像と言語情報を活用し, 大規模マルチモーダルモデルの Flamingo を基盤とした災害特化型大規模マルチモーダルモデルの開発を行った. この特化型モデルは, 従来のタスク特化型機械学習モデルを複数作成するアプローチと異なり, 単一の特化型モデルを用いて, 追加の学習を行わずに災害画像のキャプション生成, 避難指示の生成, 人数カウントといった複数の災害対応タスクに柔軟に適応できる汎用性を有する. この柔軟性により, 災害時の複雑なタスク群に対して迅速かつ高精度な対応が可能である. そのため, この特化型モデルにより, データ駆動型の災害対応の飛躍的な進展が期待される.

1 導入

機械学習技術の飛躍的な発展に伴い, 災害分野に対する機械学習の適用が進展している. 特に, ソーシャルネットワークサービス (SNS) データの解析を通じた災害情報の迅速な把握 [1, 2], 災害時の救助活動を支援する物体検出 [3], および災害予測モデル [4] の開発が注目されている. しかし, 災害時の対応は災害の事前予測, 周辺状況の把握, これらの複合要素を考慮した避難指示の提案など, 多岐にわたるタスクが内包されている. また, その一つである状況把握タスクにおいても複数のタスクに細分化される. 例えば, 周辺の建造物の損傷度, 負傷者や避難者など多様な情報を迅速に理解するタスクを扱う必要がある. このような様々なタスクに細分化される状況下で, それぞれ個別の機械学習モデルを学習させ, そのタスクに特化した機械学習モデルを作成することは非現実的である. ゆえに, 災害状況下では限られた情報, または事前情報のない状況下でも, ある程度広範囲のタスクを処理できる汎用性の高い機械学習モデルの開発が重要である.

近年, 機械学習分野でも同様な課題に直面し, その課題点を解決するために画像, 音声, テキスト等のマルチモーダル情報を統合的に処理し, 少数の学習データまたは学習データを用いずにある程度広範囲のタスクに柔軟に対応できる方法が提案されてきた. このように少数の学習データを用いて広範囲のタスクに対応する方法は, “few-shot 学習”, 学習データを用いずに対応する方法は, “zero-shot 学習” と呼ばれる. そして, このように広範囲のタスクに適用可能な機械学習モデルは, “基盤モデル” と呼ばれている. そして, これらの技術革新は, 医学 [5], 教育 [6] や法学 [7] などの他分野にも大きな影響を及ぼし, 分野特化型の大規模マルチモーダルモデルの開発を促している. しかし, 災害特化型の大規模マルチモーダルモデルは現段階で開発されていない.

そこで本研究では, オープンソースの災害関連画像とその画像に関連する言語情報を活用し, 現在さまざまな分野に応用されている大規模マルチモーダルモデル, Flamingo [8] を基盤として, 災害特化型マルチモーダルモデルを開発した. そして, 単一の特化型モデルを用いて, 災害画像のキャプション生成, 避難指示の生成, 人数のカウント等の複数のタスクに対して柔軟に対応できることを確認した. また, 特化型モデルの柔軟性により, 人数カウントタスクから派生した災害画像から負傷者数推定等にも追加学習なしで適用できることがわかった. 従来のタスク特化型の機械学習モデルの場合, このような柔軟な人数カウントを実現するためには追加の fine-tuning を要したり, 最悪の場合スクラッチから機械学習モデルを学習させる必要がある. しかし, 迅速な対応が求められる災害現場では, この追加の学習時間はボトルネックとなりうる. そのため, 我々が提案する災害特化型大規模マルチモーダルモデルのように高速かつ柔軟なモデルは, データ駆動型の迅速かつ高精度な災害時の対応能力を飛躍的に発展させ, 災害被害の大幅な軽減に直結することが期待される.

[1] ichikawa-yuma1@g.ecc.u-tokyo.ac.jp, ichikawa.yuma@fujitsu.com
[2] ryosuke.yano@tokyo-dr.co.jp

本論文では、第二章において本研究で用いる主要な大規模マルチモーダルモデルについて簡単に解説する. 第三章では, 実験に使用したモデルや学習データの詳細を示し, 第四章では, 数値実験の結果を既存の方法と比較し, 我々の特化型モデルの効果を示す. 最後に, 第五章で本論文の結論と今後の展望について述べる.

2 背景技術

本章では, 最初にマルチモーダル情報処理技術の皮切りとなった Contrastive Language-Image Pre-training (CLIP) [9] について簡単に説明する. そして, CLIP とは異なるアプローチを用いて, その適用可能性を大幅に向上させた大規模マルチモーダルモデル Flamingo [8] について説明する.

2.1 マルチモーダル基盤モデル CLIP

本章では, マルチモーダルモデル情報処理技術の発展の皮切りとなった CLIP [9] について簡単に説明する. CLIP の目的は, 画像と言語のペアとなった大規模学習データセットを用いて, 画像と言語に共通する “良い表現” を獲得し, 未知の画像の分類タスクを zero-shot または few-shot 学習により対応することである. ここで目的となる “良い表現” とは, 類似した画像同士および言語同士は類似した表現となり, また類似した画像と言語も類似した表現となることが要求される. 本章では, 画像を Image, 言語を Text と表記し, 学習データセットを $\{(\mathrm{Image}_\mu, \mathrm{Text}_\mu)\}_{\mu=1}^{P},\ P \in \mathbb{N}$ と定義する.

2.1.1 CLIP の定式化

CLIP では, 良い共通表現を獲得するために, データ同士の関係性から潜在表現を獲得する対照学習 (Contrastive Learning) という枠組みを用いる. この方法は, 教師あり学習と異なり, 画像や言語等の入力データに対する正解ラベルが不要であることから自己教師あり学習と呼ばれる手法の一つとして位置付けられる. 具体的に CLIP の対照学習では, 対応する画像と言語の組 $(\mathrm{Image}_\mu, \mathrm{Text}_\mu)$ の P 組の集合を正例データ集合, 対応しない画像と言語の組 $(\mathrm{Image}_\mu, \mathrm{Text}_\nu \mid \nu \neq \mu)$ の $P(P-1)$ 組の集合を負例データ集合を準備する. そして, 画像, 言語から M 次元特徴ベクトルを抽出する学習可能なエンコーダーをそれぞれ $I_\theta(\mathrm{Image}_\mu)$, $T_\phi(\mathrm{Text}_\mu)$ を用いてそれぞれの特徴ベクトルを獲得する. ここで, θ, ϕ は学習可能なパラメータである. さらに, その特徴表現同士の類似度スコアを次の重み付きコサイン類似度として定義する.

$$s_{\mu\nu}(\mathrm{Image}_\mu, \mathrm{Text}_\nu; \theta, \phi) = \cos\left(I_\theta(\mathrm{Image}_\mu; \theta), T_\phi(\mathrm{Text}_\mu)\right) e^t,\ \forall \mu, \nu,$$

ここで, $t \in \mathbb{R}$ は学習パラメータである. CLIP では, 重み付きコサイン類似度が用いられるが, それ以外のさまざまな類似度が提案されている. そして, CLIP は, この類似度スコアを基に, 各画像に対して対応する言語を特定する多値分類タスクと各言語から対応する画像を特定する多値分類タスクにより, 画像と言語のエンコーダ I_θ と T_ϕ を学習する. 具体的には, 画像 Image_μ に対して正解ラベルは, 対応する組の μ 番目のテキスト Text_μ となる. そのため, 正解を特徴づける one-hot ベクトル $\boldsymbol{y}^{I,\mu} = (0, \ldots, 0, 1, 0, \ldots, 0)$ と μ 行目の類似度スコアをソフトマックス関数により次のように規格化したベクトルを定義する.

$$\boldsymbol{p}^{I,\mu} = (p_1^{I,\mu}, \ldots p_i^{I,\mu}, \ldots, p_N^{I,\mu}) \in [0,1]^N,\ \ p_i^{I,\mu} = \frac{e^{s_{\mu i}}}{\sum_{i=1}^{N} e^{s_{\mu i}}}\ \ \forall i = 1, \ldots, N.$$

そして, 次のソフトマックス交差エントロピー最小化により, 画像 Image_μ に対して, 対応する組の μ 番目のテキスト Text_μ を特定する分類問題を解く.

$$\mathcal{L}^{I,\mu} = -\sum_{\nu=1}^{P} y_\nu^{I,\mu} \log p_\nu^{T,\mu}.$$

また, 各言語 Text_μ に対しても同様に, 対応する画像が 1 となる one-hot ベクトル $\boldsymbol{y}^{T,\mu} = (0, \ldots, 1, \ldots, 0)$ と μ 列目の類似度スコアをソフトマックス関数により規格化したベクトルを次のように定義する.

$$\boldsymbol{p}^{T,\mu} = (p_1^{T,\mu}, \ldots, p_i^{T,\mu}, \ldots, p_\mu^{T,N}) \in [0,1]^N, \quad p_i^{T,\mu} = \frac{e^{s_{\mu i}}}{\sum_{i=1}^N e^{s_{\mu i}}}.$$

そして, 画像の場合と同様に次のソフトマックス交差エントロピー最小化により, 言語 Text_μ に対して, 対応する組の μ 番目の画像 Image_μ を特定する分類問題を考える.

$$\mathcal{L}^{T,\mu} = -\sum_{\nu=1}^{N} y_\nu^{T,\mu} \log p_\nu^{T,\mu}.$$

最終的に, CLIP はこれらを全ての画像と言語に対して計算した次の損失関数を最小化することが学習される.

$$\mathcal{L} = \frac{1}{2}\left(\sum_{\mu=1}^{N} \mathcal{L}^{I,\mu} + \sum_{\mu=1}^{N} \mathcal{L}^{T,\mu}\right). \tag{1}$$

具体的に, CLIP では Wikipedia から抽出した画像と言語の組の 4 億個のデータを含む WIT を用いて, 式 (1) を最小化することで学習を行っている.

2.1.2 CLIP の推論

CLIP で未知画像 Image^* を分類する場合, 予めラベルの候補集合 $\{\mathrm{Text}_\mu\}_{\mu=1}^P$ を準備する. そして, 未知画像を画像エンコーダに入力し, 特徴ベクトル $I_\theta(\mathrm{Image}^*)$ と, それぞれの候補ラベルを言語エンコーダに入力した特徴ベクトル $\{T_\phi(\mathrm{Text}_\mu)\}_{\mu=1}^P$ を取得し, 未知画像の特徴ベクトルに対する各言語特徴ベクトルの類似度スコア $\{s_{*\mu}\}_{\mu=1}^P$ を計算し, スコアが最大となる Text^{μ^*}, $\mu^* = \mathrm{argmax}_\mu\{s_{*\mu}\}_{\mu=1}^P$ を分類結果とする. この方法の最大の利点は, 従来の画像分類モデルと異なり, 任意のラベルを事前に準備し, モデルの fine-tuning なしで推論が可能である点である. この学習手法は zero-shot 学習に分類される. そして, このように幅広い分類問題等に適用可能なモデルを “基盤モデル” と呼ぶ. CLIP の技術に動機づけられ, 音声-言語 [10], 動画-言語 [11], 画像-音声-言語 [12] の CLIP ベースの基盤モデルが提案されている. また, 対照学習により得られる画像と言語の共通表現を取得可能なエンコーダは後述の Flamingo 等の近年の大規模マルチモーダルモデルに利用される.

2.2 大規模マルチモーダルモデル Flamingo

対照学習に基にした基盤モデルは, Fine-tuning を行わずに zero-shot 学習または few-shot 学習により新たな分類問題に対して高性能な分類が実現可能である. しかし, 対照学習は言語との類似度スコアを学習の土台として用いるため, 分類問題のように比較的単純なユースケースへの適用に限定される. 実際, 画像等の視覚情報の理解—例えば, 視覚情報のキャプション生成や視覚情報に関係する質問に答えるタスク—が必要不可欠な広範囲なユースケースに適用する推論精度が大幅に低下することが知られている.

その問題に対して, 対照学習とは異なり事前学習された各モーダルの基盤モデルを効果的に接合するアプローチにより, この限界を緩和する方法が提案されている. 本章では, この方法の一つの先駆的な研究である DeepMind が提案した大規模画像言語生成モデル, Flamingo を紹介し, マルチモーダル情報処理に対して, 各モーダルの基盤モデルの多様性を維持したまま接合する方法論を概観する. また, Flamingo は, 基盤モデルの接合による汎用性の向上のみならず. 任意のサイズの画像や動画をシームレスに処理する能力を有する. この技術は, 画像や動画の視覚情報に限定されるアプローチではなく, 一般の可変長データを扱う上で必要不可欠となる技術的な革新である. この方法に関しても簡単に説明する.

Flamingo による Few-shot 学習の結果を図 1 に示す. Flamingo は, たった 32 個の学習データを用いて 16 個の画像および動画を理解するタスクのうち 6 個のタスクに対して既存のタスク特化型機械学

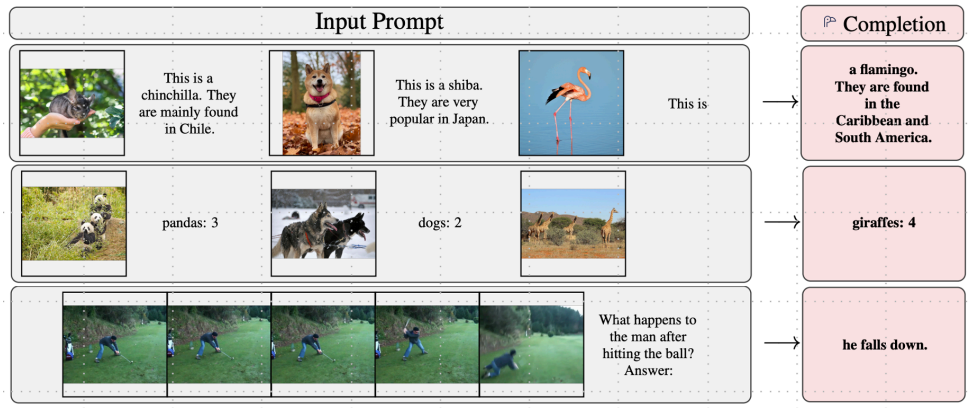


図 1: Flamingo-80B の入出力例 [8]. Flamingo は, 数ショットのプロンプトで様々な画像動画タスクに対応可能.

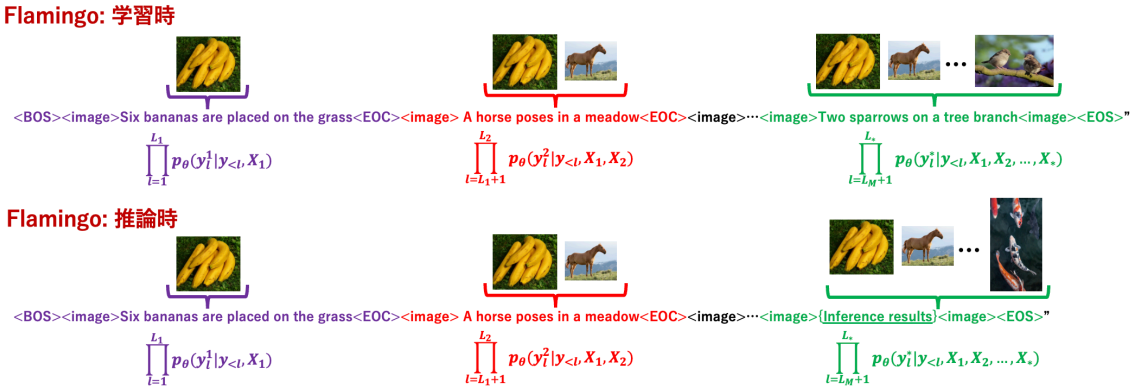


図 2: Flamingo の統計モデルの概念図. 上が学習時, 下が推論時の様子を表す.

習モデルの結果を上回る性能を実現した. 以降, Flamingo の統計モデルとその統計モデルのネットワーク構造および学習方法について詳しく説明する.

2.2.1 Flamingo の統計モデル

Flamingo は, 画像または動画条件付きテキスト自己回帰モデルである. 以降, 画像または動画をまとめて視覚情報と呼ぶことにする. 形式的に確率分布は, M 個の視覚情報 $\{X^m\}_{m=1}^M$ と言語 $\{\boldsymbol{y}^m \in \mathbb{R}^{L_m}$ のペア $\mathcal{D}_{\text{Few}} = \{(X^m, \boldsymbol{y}^m)\}_{m=1}^M$ および, 推論対象の視覚情報 X^* と言語 $\boldsymbol{y}^* \in \mathbb{R}^L$ からなる言語列 $\boldsymbol{y} = (\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^M, \boldsymbol{y}^*)$ と視覚情報列 $\boldsymbol{X} = (X^1, \ldots X^M, X^*)$ に対して, 次のように定義される.

$$p_\theta(\boldsymbol{y}|\boldsymbol{X}) = \prod_{l=1}^{L_1} p_\theta(y_l^1|\boldsymbol{y}_{<l}, X^1) \prod_{l=L_1+1}^{L_2} p_\theta(y_l^2|\boldsymbol{y}_{<l}, X^1, X^2) \cdots \prod_{l=L_N+1}^{L^*} p_\theta(y_l^*|\boldsymbol{y}_{<l}, X^1, \ldots, X^*), \quad (2)$$

ここで, θ は Flamingo の学習可能なパラメータである. また, $\boldsymbol{y}_{<l} \triangleq (y_1^1, \ldots, y_{l-1}^k \mid L^{k-1} < \sum_{l'=1}^{l} 1 \leq L^k)$ と定義する. そして, この確率分布のパラメータ θ は最尤推定に基づき対数尤度の最大化により学習される. 先行研究では, $M = 5$ の場合の対数尤度最大化により学習が行われている. この統計モデルの概念図を図 2 に示す. さらに, 先行研究では K 種類の学習データセットに対して, 次の重みづけ対数尤度を用いて学習を行っている.

$$\mathcal{L}(\theta) = \sum_{k=1}^{K} \lambda_k \mathbb{E}_{(X,y)\sim\mathcal{D}_k} \left[-\log p_\theta(\boldsymbol{y}|\boldsymbol{X}) \right],$$

ここで, $\mathcal{D}_k$ は k 番目の学習データセット, λ_k はその学習データセットに対する重みを表す. また, $\mathcal{E}_{(X,y)\sim\mathcal{D}_k}[\cdot]$ は, 学習データ $\mathcal{D}_k$ に関する経験平均を表す.

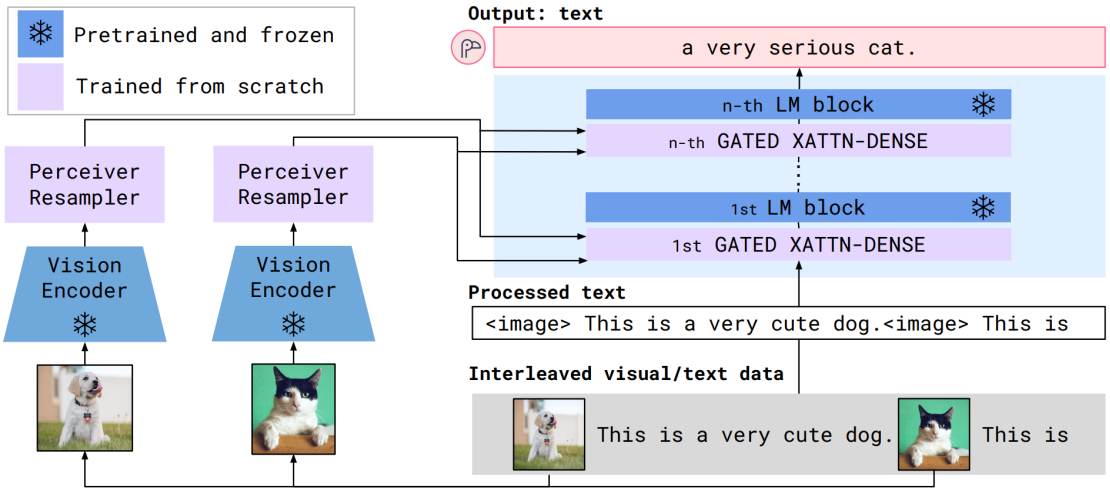


図 3: Flamingo のネットワーク構造の概念図 [8]

そして, 先行研究では, 18 億個の画像と Alt-Text からなる ALIGN [13], 3 億 1200 万個の画像と比較的長い言語からなる LTIP, 2700 万個の短い動画とその動画のディスクリプションからなる VTP と 4300 万個の Web 上のデータから作成した M3W を用いて, それぞれの λ_k を調節して学習を行なった. より詳しい重みの調整方法に関しては, 先行研究 [8] の Appendix B を参照していただきたい.

2.2.2 Flamingo の学習手法

2.2.1 章で述べたように, 形式的には Flamingo は, 式 eq:prob-model-flamingo の学習パラメータ θ を最尤推定により決定することで学習がなされる. しかし, 単に適当なモデルを用いたとしても zero-shot 学習または few-shot 学習を行うことは不可能である. そこで, Flamingo は既存の画像および言語に特化した基盤モデルのパラメータを固定して, その条件付けを支配する接合部のみを学習する方法を採用する. その概念図を図 3 に示す. 図の `Vision Encoder` は画像の基盤モデルの画像エンコーダー (画像の特徴量抽出機) を表し, `LM block` は言語の基盤モデルの一部を表し, `1st LM block` から `n-th LM block` 全体が言語の基盤モデルとなる. これらの基盤モデルは, 式 (2) の θ に含まれず, θ の学習時には固定される定数である [3]. そして, 図 3 の `Perceiver Resampler` と `GATED XATTN-DENSE` が画像基盤モデルの画像エンコーダと言語基盤モデルの接合部の役割を担う. そのため, その接合部のみが学習可能なパラメータ θ に特徴付けられ, 式 (2) の対数尤度最大化により学習される. このように, 既に大量の画像や言語で学習されている基盤モデルの学習を再度行わず, その接合部のみを学習することで大幅な高速化を実現している. さらに, 基盤モデルを固定することで, その基盤モデルが既に獲得している多様性を維持することが可能である. 実際, 深層モデルでは “破滅的忘却 [14]” と呼ばれ, 次のタスクを学習するとそれ以前のタスクを大幅に忘却することが知られており, その問題を基盤モデルの固定により回避している. 実際, 先行研究 [8] では, 数値的に画像, 言語の基盤モデルも学習可能なパラメータ θ に含めると性能が大幅に悪化することを確認している. 以降, 接合部の `Perceiver Resampler` と `GATED XATTN-DENSE` について簡単に説明する.

Perceiver Resampler. `Perceiver Resampler` は, Jaegle らによって導入された Perceiver model [15] を参考に構成されている. Perceiver Resampler は, さまざまな画像や動画の `Vision Encoder` によって出力される可変長の特徴ベクトルを固定長のベクトルに変換するモデルである. この Perceiver Resampler は直感的には, 視覚情報を理解し, 言語を生成するタスクにおいて, 真に重要な可変長の特徴ベクトルの一部を “Attention 機構” と呼ばれる枠組みでさらに圧縮し, 固定長のベクトルに変換

[3] 先行研究 [8] では, 学習時に固定することを “freeze” という表現で記述していることに注意.

図 4: 本研究で用いた学習データセットの一部. 左が消防防災科学センターが公開している災害画像データベース. 右が Kaggle で公開されている Disaster Image Dataset である.

する技術である. より具体的なネットワークおよび実装方法は, 先行研究 [8] の Appendix A を参照していただきたい.

Gated Cross-attention. 図 3 にあるように, `GATED XATTN-DENSE` は, 言語基盤モデルの途中に挿入され, `Perceiver Resampler` の出力の固定長ベクトルを入力とし, 次の `LM block` に視覚情報の情報を伝達する役割を担う. 具体的には, `Perceiver Resampler` の出力の固定長ベクトルを $\hat{X}$ とし, 一つ前の `(i-1)-th LM block` の出力を $\hat{y}^{i-1}$ とすると次の出力 $\hat{h}^{i-1}$ を次の `I-th LM block` に情報伝達する.

$$\hat{h}^{i-1} = \hat{y}^{i-1} + \tanh(\alpha) f(\hat{y}^{i-1}, \hat{X}; \theta), \tag{3}$$

ここで, θ と $\alpha \in \mathbb{R}$ は学習可能なパラメータである. また, 式 2 の対数尤度を上げるように, f は言語情報 $\hat{y}^{l-1}$ に関係する視覚情報 X を抽出するような Attention 機構が利用されている. 具体的な関数系と実装方法は, 先行研究 [8] の 2 章を参照していただきたい. また, 式 (3) の第二項の $\tanh(\alpha)$ は, $\alpha = 0$ のときは, 既存の言語基盤モデルを復元することが可能である. この α は言語モデルが視覚情報を利用する程度を表し, この α を確認することで視覚情報の条件付けの効果がどの程度取り入れられているかを確認することが可能である. 先行研究では, 学習の初期化時に $\alpha = 0$ を用いると学習の安定性および性能が向上することを報告している.

3 実験設定

本研究では, Flamingo を学習させるために, 消防防災科学センターが公開しているタイトル付きの災害画像データベース (`http://www.saigaichousa-db-isad.jp/drsdb_photo/photoSearchResult.do`) と Kaggle で公開されている Disaster Images Dataset (`https://www.kaggle.com/datasets/varpit94/disaster-images-dataset`) を使用する. それぞれの画像のデータセットの一部を図 4 に示す. そして, 本研究では公開されている open flamingo [16][4]の Open Flamingo-9B を用いて数値実験を行った. この Open Flamingo-9B モデルは, 図 3 の構成要素に次のモデルを使用した.

- `Vision Encoder`: CLIP の事前学習済み Vision Transformer (1B)
- `LM Model`: LLaMA (7B) [17]
- `Perceiver Resampler` (1B)
- `GATED XATTN-DENSE Layer`, 4 個の `LM block` ごとに設定

[4] `https://github.com/mlfoundations/open_flamingo`

図 5: Flamingo によるキャプション生成タスクと避難指示生成タスクの推論結果の具体例. 赤字で書かれた結果が Flamingo によって生成された文章. 右はキャプション生成タスクの例で, 左は避難指示生成タスクの few-shot の例と避難指示の生成結果である.

現在, open flamingo[5] [16] は, 他の言語モデルも使用可能だが, 本研究では, もっとも広く用いられている LLaMA (7B) [17] を使用する.

4 実験結果

4.1 災害画像のキャプション生成タスク

まず, 最も基本的な数値実験としてタイトル付きの災害画像データベースを用いて, 災害画像のキャプション生成タスクを行った. タイトル付き画像であるため評価指標として, Flamingo-9B によって生成されたキャプションとタイトルをそれぞれ事前学習済みのテキストエンコーダ [18] を用いて特徴ベクトルに変換し, その特徴ベクトル同士のコサイン類似度を使用する. その結果, 16 個の web 上の画像を few-shot として用いるデフォルトの Flamingo-9B の場合, コサイン類似度は, 0.78 ± 0.0023 であった. ここで, 誤差は Flamingo の 5 回の推論結果の標準偏差を表す. 一方, Flamingo-9B を災害画像データベースを用いて fine-tuning を行い, 16 個の few-shot を用いて学習した結果は, 0.92 ± 0.0033 となり, デフォルトの Flamingo-9B の性能を上回った. この結果から, 僅かなデータセットであるが, Flamigo を災害分野に特化することで性能向上が見込める可能性が示唆された.

4.2 災害画像の避難指示生成タスク

次に災害画像から避難指示を生成するタスクに対して, Flamingo-9B を適用した結果を示す. 避難指示タスクは, Fine-Tuning した Flamingo を用いても, 定性的にあまり有益な避難指示を生成することは困難であった. しかし, 図 4.2 のように, 画像に加えて, "A flood occurs" のような災害状況を付加することである程度妥当な避難指示を生成することが可能となった. しかし, その避難指示が妥当性の判断は, 正解の避難指示の情報がないためどの程度妥当な避難指示を生成できているかを定量的に評価することは困難である. この定量的な比較を防災に関する専門家の方々と議論して評価することは今後の展望である.

4.3 災害画像の人数カウントタスク

最後に, 災害画像を入力として画像内の人数をカウントするタスクを行った. この数値実験では, Kaggle の Disaster Image Dataset の human damage の画像上の人数を目測で事前にカウントしたデータを作成し, その人数を正解データとして使用した[6]. そして, Flamingo が出力する人数と正解

[5] https://github.com/mlfoundations/open_flamingo
[6] 目測で数えた人数のデータに関しては, https://github.com/Yuma-Ichikawa/DisasterLMM/tree/main を参照していただきたい.

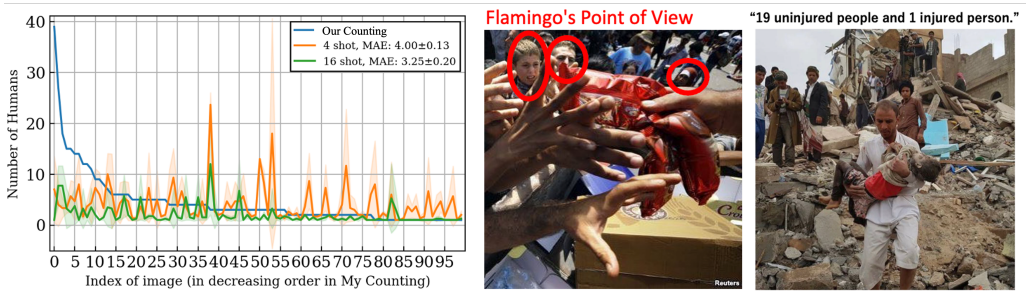


図 6: (左) 縦軸: 人数, 横軸: 推論対象画像の ID (目測の人数が大きい順), MAE: 予測誤差の絶対値の平均, 誤差は Flamingo の 5 回の推論の標準偏差, (中央) Flamingo が人数カウントタスクで認知している領域の例, (右) 怪我を負っている人数と怪我をしていない人数をカウントするタスクに対して, Flamingo を適用した結果.

の人数の平均絶対誤差 (MAE) を用いて評価を行った. Flamingo に画像と正解の人数の組みを 4shot 与えた場合の MAE は, 4.00 ± 0.13 であった. ここで, 誤差は Flamingo の 5 回の推論の標準偏差を表す. 一方, 16shot 与えた場合の MAE は, 3.25 ± 0.20 であり, few-shot の数を増やすことで性能が向上した. また, Flamingo が各画像に対して推論を行った結果と正解データの違いを図 6 に示す. この結果から, 人数が多い領域の推論精度が低いことがわかった. その一つの原因として, 我々の目測結果は腕や脚の一部から人数を推論することが可能だが, Flamingo は, 図 6 の中央の図のように顔の情報から人数をカウントしている傾向があることがわかった. この課題点は, データ拡張により, 腕や脚の一部から人数をカウントするような学習データセットを作成し, Flamingo を再学習することで改善される可能性がある.
また, 単純な人数カウントに加えて, 図 6 の右図にあるような怪我人と怪我人でない人数を数えるタスクに関してもある程度の精度で人数をカウントできることがわかった. このように単純に few-shot で柔軟な人数カウントができることが Flamingo をはじめとする大規模マルチモーダルモデルの強みである.

5 終わりに

本論文では, 大規模マルチモーダルモデルの学習方法について簡単に概観し, その後 Flamingo を土台に災害特化型の大規模マルチモーダルモデルの開発を行った. このモデルは, 従来の機械学習の枠組みとは異なり, 各タスクにそれぞれの特化型機械学習モデルを用いずに, 単一の大規模マルチモーダルモデルを用いて, 災害時の避難指示生成や人数カウントなど, 幅広い災害関連タスクにおいてある程度高い推論が可能である. このような汎用性は, 迅速かつ精確な判断が要求される災害時対応において, 重要な結果である. また, 導入で説明した基盤モデルを接合して大規模マルチモーダルモデル学習するアプローチは, 高い汎用性を持つ. そのため, 災害状況を数値化する観測データや数理モデルを含む様々な情報を効率的に統合し, より災害に特化した大規模マルチモーダルモデルの開発が重要な方向性であり, 今後の展望である.

References

[1] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. *IEEE intelligent systems*, Vol. 27, No. 06, pp. 52–59, 2012.

[2] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. Extracting situational information from microblogs during disaster events: a classification-

summarization approach. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 583–592, 2015.

[3] Subrahmanyam Vaddi. *Efficient object detection model for real-time UAV applications*. PhD thesis, Iowa State University, 2019.

[4] M Apriani, SK Wijaya, et al. Earthquake magnitude estimation based on machine learning: Application to earthquake early warning system. In *Journal of Physics: Conference Series*, Vol. 1951, p. 012057. IOP Publishing, 2021.

[5] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.

[6] Kamil Malinka, Martin Peresíni, Anton Firc, Ondrej Hujnák, and Filip Janus. On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, pp. 47–53, 2023.

[7] Dietrich Trautmann, Alina Petrova, and Frank Schilder. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*, 2022.

[8] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, Vol. 35, pp. 23716–23736, 2022.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[11] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, Vol. 508, pp. 293–304, 2022.

[12] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 20020–20029, 2022.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

[14] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, Vol. 24, pp. 109–165. Elsevier, 1989.

[15] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

[16] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023.

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.