大規模言語モデルを活用した Vison-Language Modelによる詳細画像分類

B4　岩村班　川越　壮

目次

- 研究背景
- 提案手法
- 実験
- まとめ・今後の方針

目次

研究背景

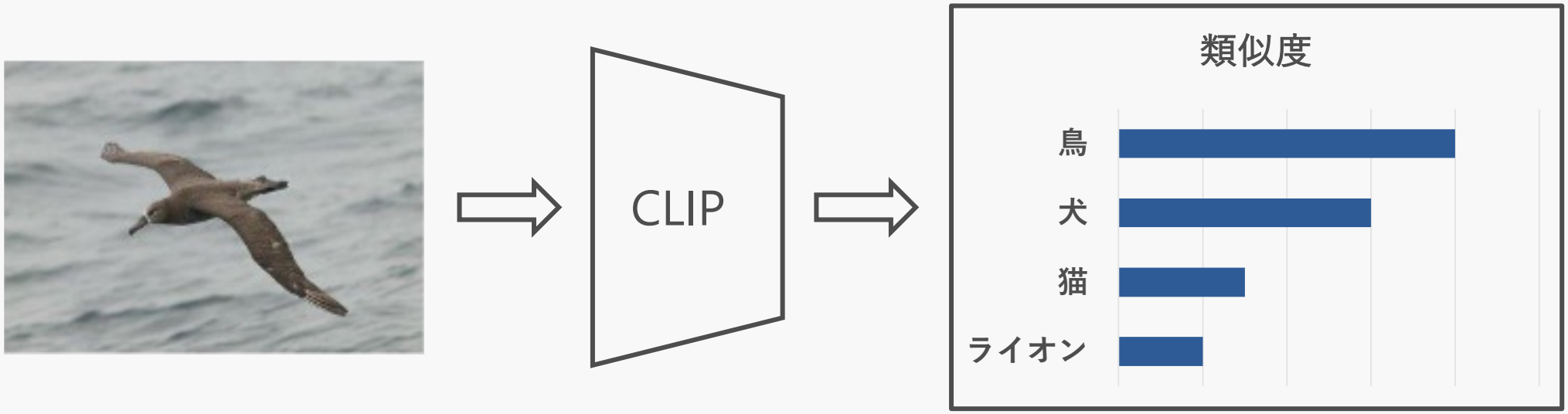
- Vision-Language Model(VLM)による画像分類

  画像処理と自然言語処理を融合したモデル

  - CLIP[1]

    画像とテキストを同一の特徴空間に移し，ペアの特徴量が近くなるように学習

    画像とクラス名の類似度からクラスを推論



[1] Alec Radford, et al. Learning transferable visual models from natural language supervision, 2021

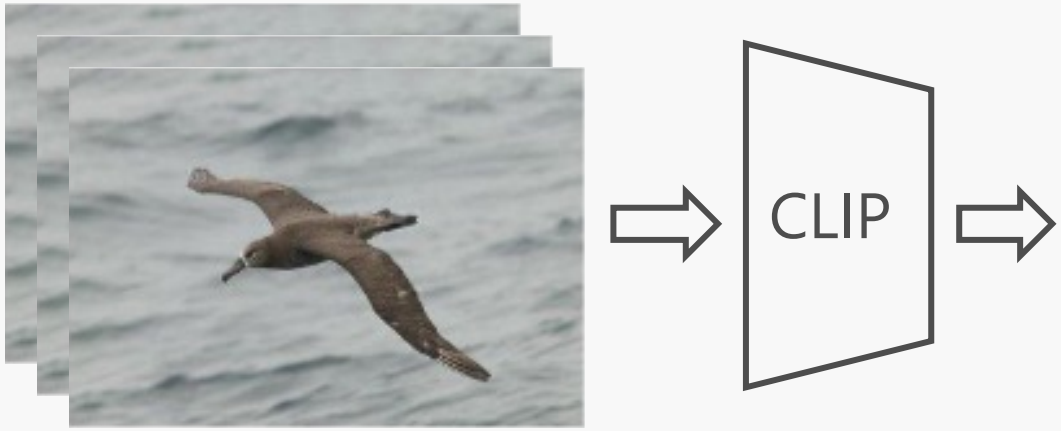
研究背景

- Vision-Language Model(VLM)の課題

  詳細な画像分類タスクにおいてのパフォーマンスが悪い

  →**クラス名と視覚属性を付帯したクラス名**で推論精度はほとんど変わらない[2]

  例：CUB（様々な鳥類を含むデータセット）の場合



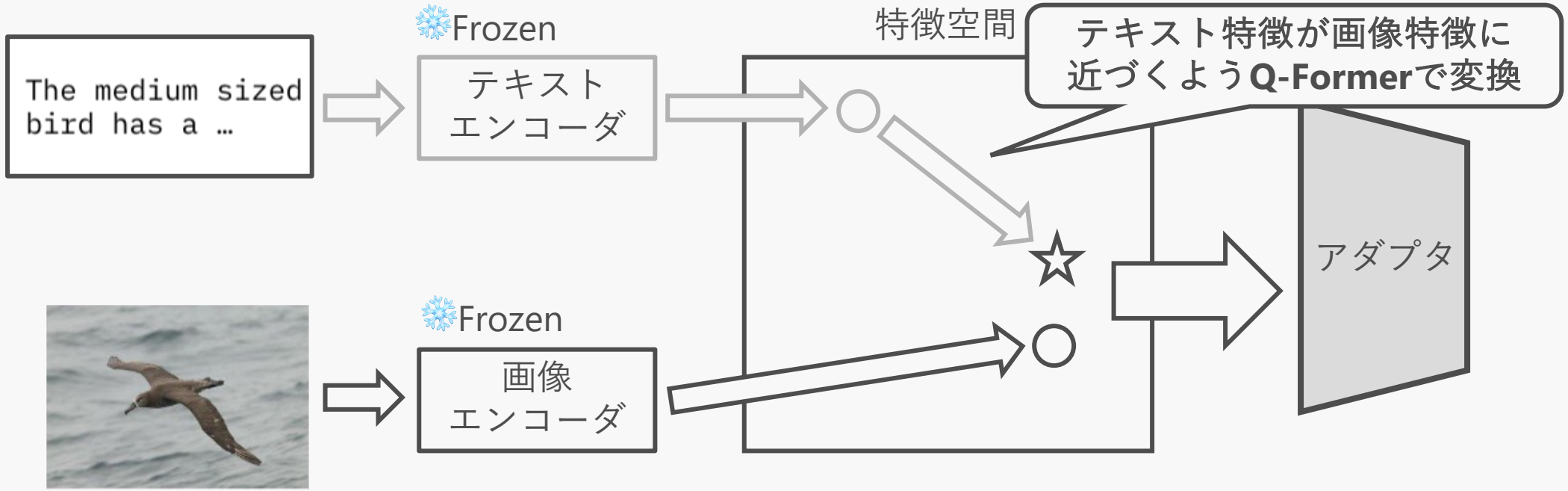
クラス名 : **50.5%**
（例 Black Footed Albatross）
クラス名＋視覚属性 : **50.7%**
（例 Black Footed Albatross with a dark gray body…）

[2] Oindrila Saha, et al. Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions. In CVPR, 2024.

既存研究

平野の研究[1]

- **二段階の学習でテキストを画像のように扱い学習することで詳細画像分類にアプローチ**
  1. Q-Former[2]を用いてテキスト特徴から画像特徴への変換を学習
  2. 学習したQ-Formerと説明文を用いてアダプタ（小規模ネットワーク）を学習



[1]平野甫．テキスト特徴量から画像特徴量への変換による詳細画像分類．大阪府立大学大学院工学研究科電気・情報系専攻知能情報工学分野修士学位論文，2024.
[2] Junnan Li, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

既存研究の結果

- 画像とテキストの割合を変えて学習をした結果

割合(画像, テキスト)	画像のみ	画像＋テキスト
(100 ,0)	84.80	—
(90, 10)	84.67	**85.05**
(80, 20)	83.55	**84.55**
(70, 30)	82.55	**84.16**
(60, 40)	80.48	**83.33**
(50, 50)	79.81	**82.52**
(40, 60)	77.82	**80.84**
(30, 70)	**73.33**	67.50
(20, 80)	**67.24**	61.25
(10, 90)	**53.92**	49.12
(0, 100)	—	27.67

既存研究の課題

- 平野の研究では人の手で付けられたテキストを学習に使用

Reedら[2]が各画像に10個のテキストを作成

- the medium sized bird has a dark grey color, a black downward curved beak, and long wings.
- the bird is dark grey brown with a thick curved bill and a flat shaped tail.
- this bird is gray in color, with a large curved beak.

⋮

画像のラベル付け以上にコスト・専門的知識が必要

[2]: Scott Reed, et al. Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016

大規模言語モデル（Large Language Model）

- Web上の大量のテキストデータを学習
- カテゴリに対して詳細な説明文を指定した構造で生成可能
- APIを用いることで効率的にテキスト出力が可能

クラス名 : Black Footed Albatross

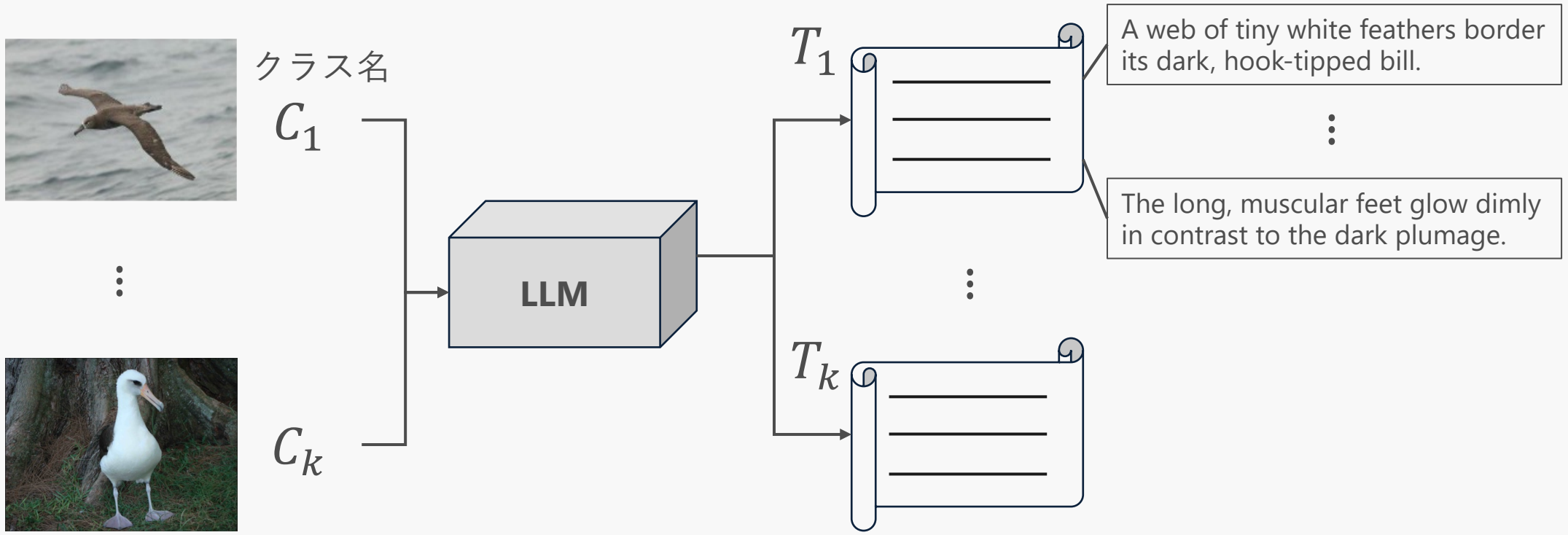
LLM

Size: It has a wingspan of about 6.5 to 7 feet (2 to 2.1 meters) and a body length of around 28 to 32 inches (71 to 81 cm).
Coloration: The plumage is predominantly dark brown to black, with the exception of some white feathers around the base of the beak and under the eyes.
Beak: It has a large, strong, black beak with a hooked tip.
Feet: The legs and feet are black, which is a key identifying feature.
Eyes: The eyes are dark, blending in with the surrounding plumage.

目次

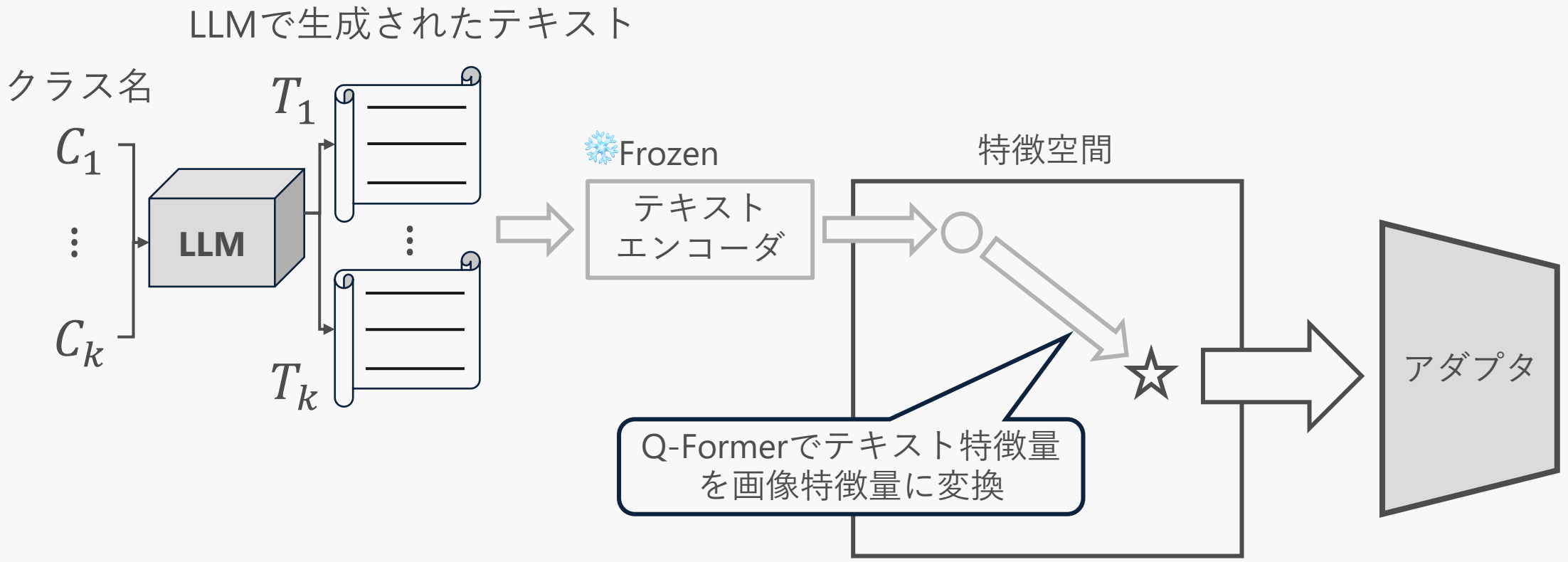
提案手法

- 大規模言語モデルを用いてクラスの視覚的情報に関するテキストを生成
- 生成させたテキストを用いてアダプタを学習



提案手法

- 学習の流れ


LLMで生成されたテキスト
クラス名
C_1
C_k
LLM
T_1
T_k
Frozen
テキスト
エンコーダ
特徴空間
Q-Formerでテキスト特徴量
を画像特徴量に変換
アダプタ

目次

実験

- 画像エンコーダ，テキストエンコーダには学習済みのCLIPを使用
- アダプタは1層の全結合層

- Q-Formerの学習　データセット : Microsoft COCO
- アダプタの学習　データセット : CUB-200-2011
- テキストを生成させるLLM : gpt-3.5-turbo

平野と同条件

- 画像割合を変化させる際に学習サンプルを全体からランダムに選択

→学習サンプルにクラスの偏りがあった可能性

→クラス単位での割合に修正

実験結果

割合(%) (画像, テキスト)	画像のみ (平野)	画像＋テキスト	
		Reedら (平野)	gpt-3.5-turbo
(100 ,0)	**85.07（84.80）**	—	—
(90, 10)	**85.26（84.67）**	83.87（85.05）	84.05
(80, 20)	**84.95（83.55）**	83.13（84.55）	82.90
(70, 30)	**84.71（82.55）**	81.57（84.16）	81.62
(60, 40)	**84.48（80.48）**	80.85（83.33）	80.01
(50, 50)	**84.23（79.81）**	77.64（82.52）	78.68
(40, 60)	**83.66（77.82）**	76.10（80.84）	76.56
(30, 70)	**83.38（73.33）**	72.05（67.50）	72.47
(20, 80)	**82.83（67.24）**	68.08（61.25）	67.24
(10, 90)	**81.88（53.92）**	57.47（49.12）	57.96
(0, 100)	—	24.12（27.67）	23.78

実験結果：テキストのみで学習

割合(%) (画像, テキスト)	画像のみ (平野)	画像＋テキスト	
		Reedら (平野)	gpt-3.5-turbo
(100 ,0)	85.07（84.80）	—	—
(90, 10)	85.26（84.67）	83.87（85.05）	84.05
(80, 20)	84.95（83.55）	83.13（84.55）	82.90
(70, 30)	84.71（82.55）	81.57（84.16）	81.62
(60, 40)	84.48（80.48）	80.85（83.33）	80.01
(50, 50)	84.23（79.81）	77.64（82.52）	78.68
(40, 60)	83.66（77.83）	76.10（80.04）	76.56
(10, 90)	81.88（55.92）	57.47（49.12）	57.96
(0, 100)	—	**24.12**（27.67）	**23.78**

テキストのみでモデルを学習した場合，学習に用いたテキストが，
人の手で作られたものと**LLMで生成したもの**とで精度の差はあまりない

実験結果：画像のみで学習

割合(%) (画像, テキスト)	画像のみ (平野)	画像＋テキスト	
		Reedら (平野)	gpt-3.5-turbo
(100 ,0)	**85.07 (84.80)**	—	—
(90, 10)	**85.26 (84.67)**	83.87 (85.05)	84.05
(80, 20)	**84.95 (83.55)**	83.13 (84.55)	82.90
(70, 30)	**84.71 (82.55)**	81.57 (84.16)	81.62
(60, 40)	**84.48 (80.48)**	80.85 (83.33)	80.01
(50, 50)	**84.23 (79.81)**	77.64 (82.52)	78.68
(40, 60)	**83.66 (77.82)**	76.10 (80.84)	76.56
(30, 70)	**83.38 (73.33)**	72.05 (67.50)	72.47
(20, 80)	**82.83 (67.24)**		
(10, 90)	**81.88 (53.92)**	57.47 (49.12)	57.96
(0, 100)	—	24.12 (27.67)	23.78

画像のみでの精度が大幅に向上

実験結果：画像とテキストで学習

割合(%) (画像, テキスト)	画像のみ (平野)	画像＋テキスト	
		Reedら (平野)	gpt-3.5-turbo
(100 ,0)	**85.07** (84.80)	—	—
(90, 10)	**85.26** (84.67)	83.87 (85.05)	84.05
(80, 20)	**84.95** (83.55)	83.13 (84.[illegible])	[illegible]
(70, 30)	**84.71** (82.55)	81.57 (84.16)	81.62
(60, 40)	**84.48** (80.48)	80.85 (83.33)	80.01
(50, 50)	**84.23** (79.81)	77.64 (82.52)	78.68
(40, 60)	**83.66** (77.82)	76.10 (80.84)	76.56
(30, 70)	**83.38** (73.33)	72.05 (67.50)	72.47
(20, 80)	**82.83** (67.24)	68.08 (61.25)	67.24
(10, 90)	**81.88** (53.92)	57.47 (49.12)	57.96
(0, 100)	—	24.12 (27.67)	23.78

テキストを学習に用いた場合の精度が**画像のみ**の精度を下回る

目次

まとめ・今後の方針

まとめ

- テキストを用いてアダプタを学習するという平野の手法において，LLMで生成させたテキストは有効
- クラスの偏りをなくしたことで，画像のみを学習に用いた場合の精度が向上し，テキストを学習に用いると精度が低下

今後の方針

- テキスト特徴量から画像特徴量への変換精度の調査

大規模言語モデルを活用した Vison-Language Modelによる詳細画像分類

B4　岩村班　川越　壮