大規模言語モデルを用いた Vison-Language ModelによるFew-shot分類

B4　DL班　川越　壮

目次

目次

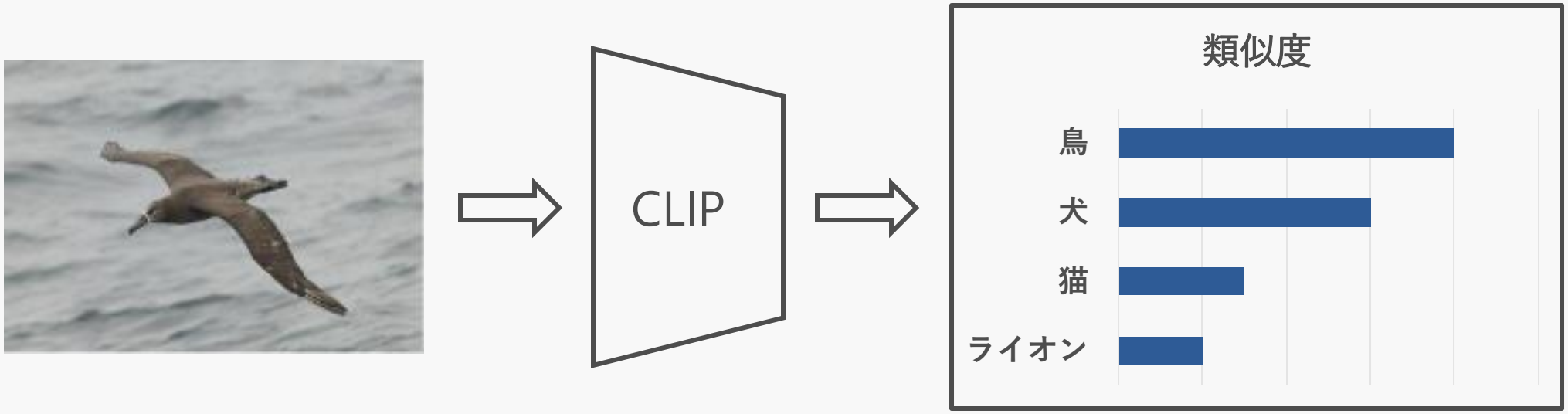
研究背景

- 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

  - Contrastive Language-Image Pre-training (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

      →汎化性能が高いため，Few-shotや教師あり学習で活用したい



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

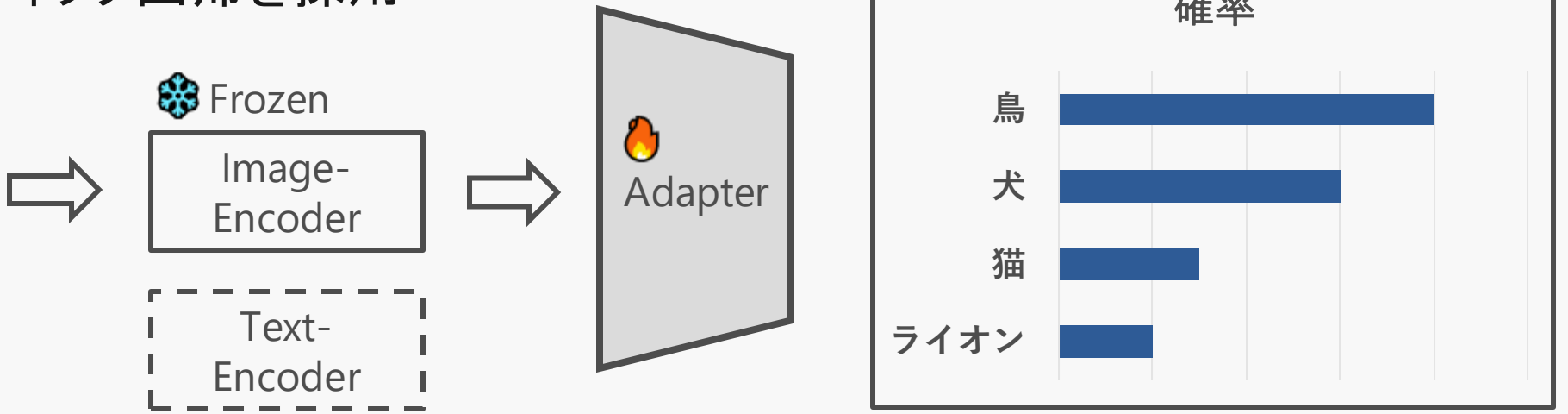
Few-shotや教師あり学習への応用

CLIPをファインチューニングする際の問題

- モデルサイズが膨大なためコストが高い
- 破滅的忘却

■ Linear-Probe[2]

- Image-Encoderの末尾にAdapter（小規模の分類器）
- ロジスティック回帰を採用

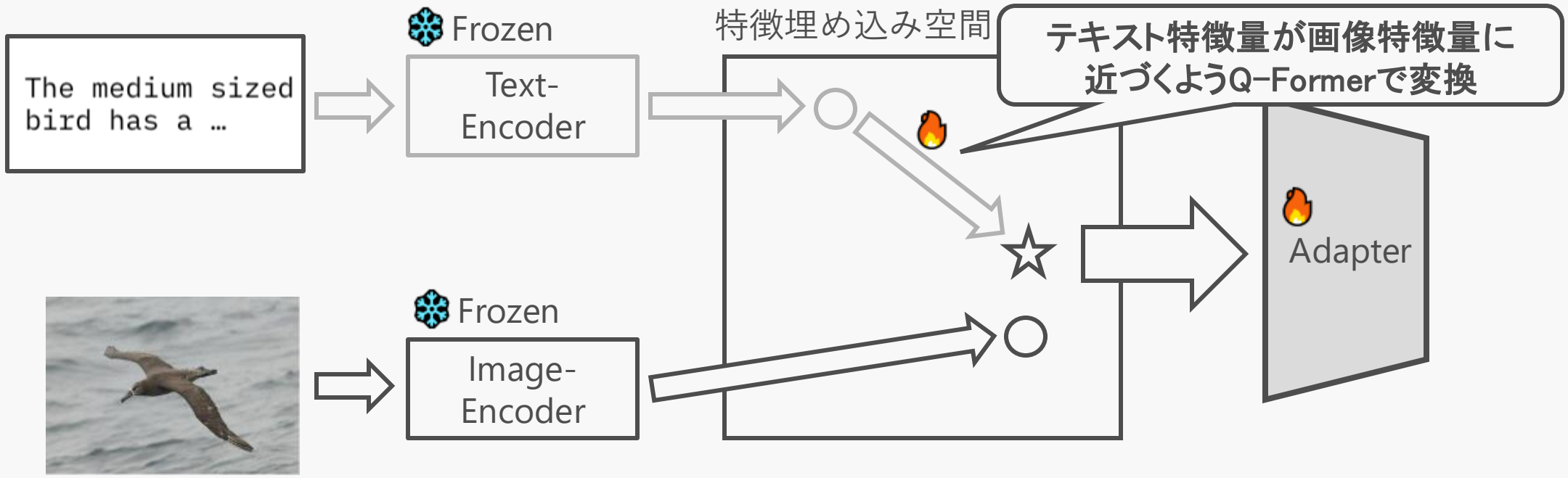


[2] Alec Radford, et al. Learning transferable visual models from natural language supervision, 2021

既存研究

平野の研究[3]

- **2段階の学習により，テキストを画像のような扱いが可能に**
  1. Q-Former[4]を用いてテキスト特徴量から画像特徴量への変換を学習
  2. 学習したQ-Formerと説明文を用いてアダプタ（ニューラルネットワーク）を学習



[3]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.
[4] Junnan Li, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

既存研究の課題

- 下流タスクのデータセットにはテキストが付いていない
- 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

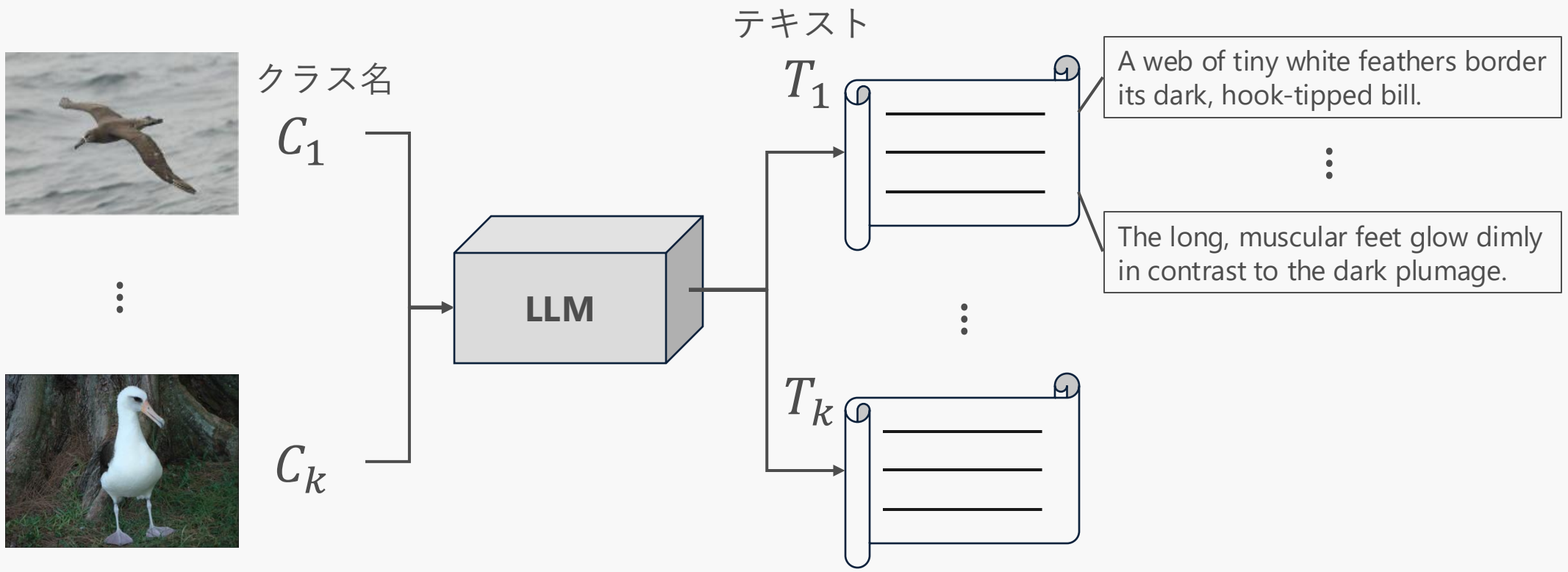
Chat GPT

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

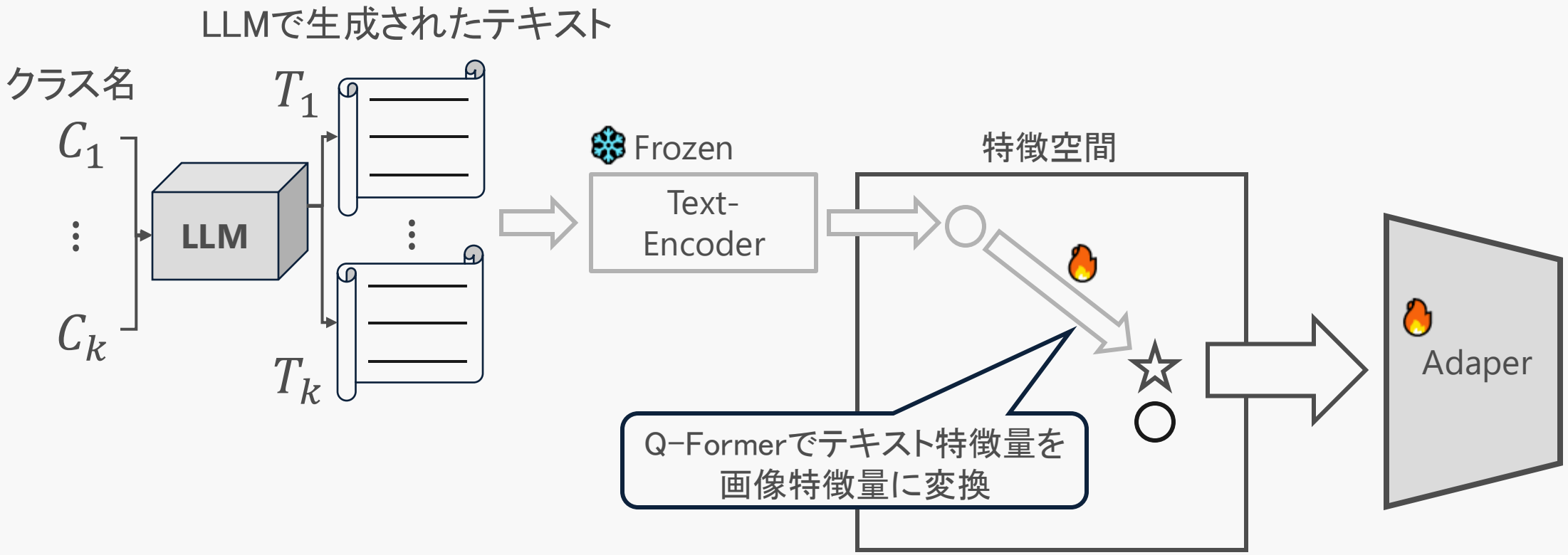
提案手法

- 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Adapter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
Text-
Encoder
特徴空間
Q-Formerでテキスト特徴量を
画像特徴量に変換
Adaper

目次

実験：Few-shot Classification

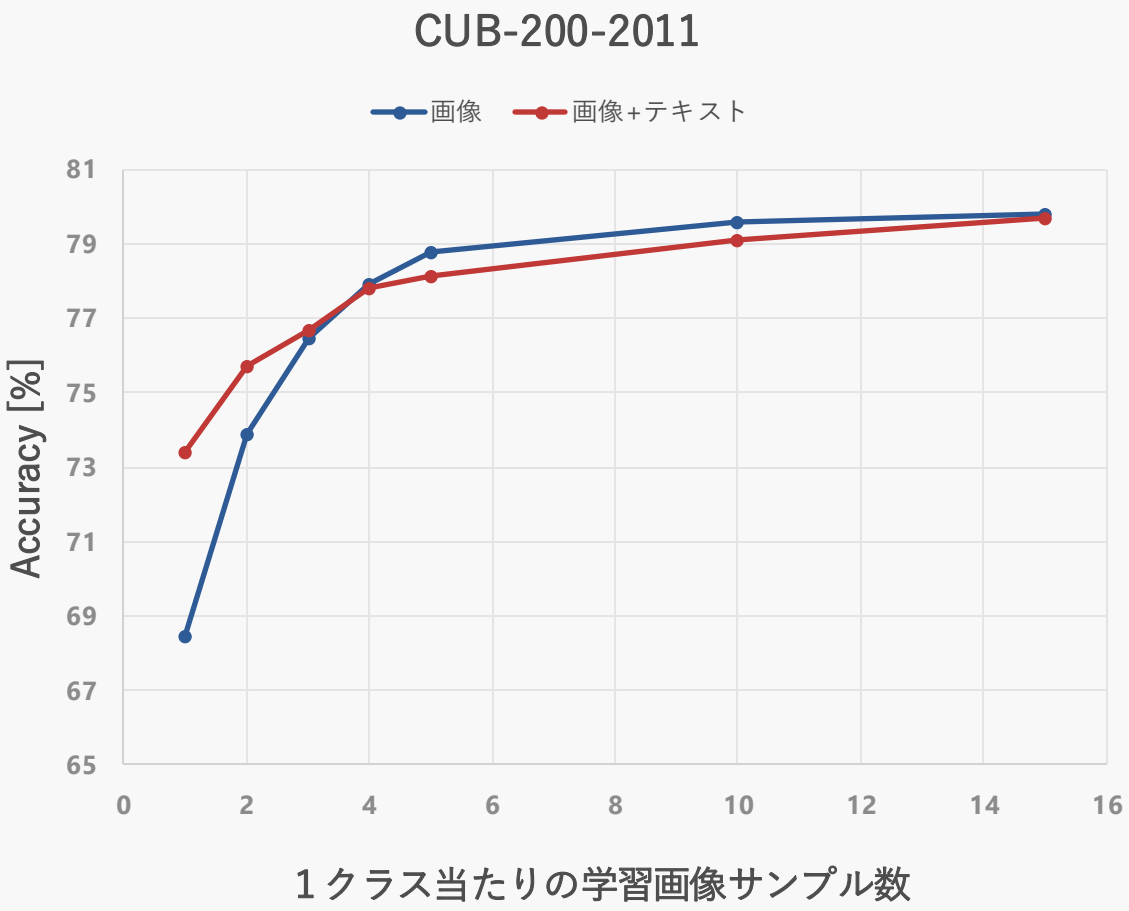
- Image-Encoder，Text-Encoderには学習済みのCLIP（ViT B/16）を使用
- Adapterは1層の全結合層
- テキストを生成させるLLM：gpt-3.5-turbo
- データセット
  - CUB-200-2011（200種類の鳥）
  - OxfordPets（100種類の動物）

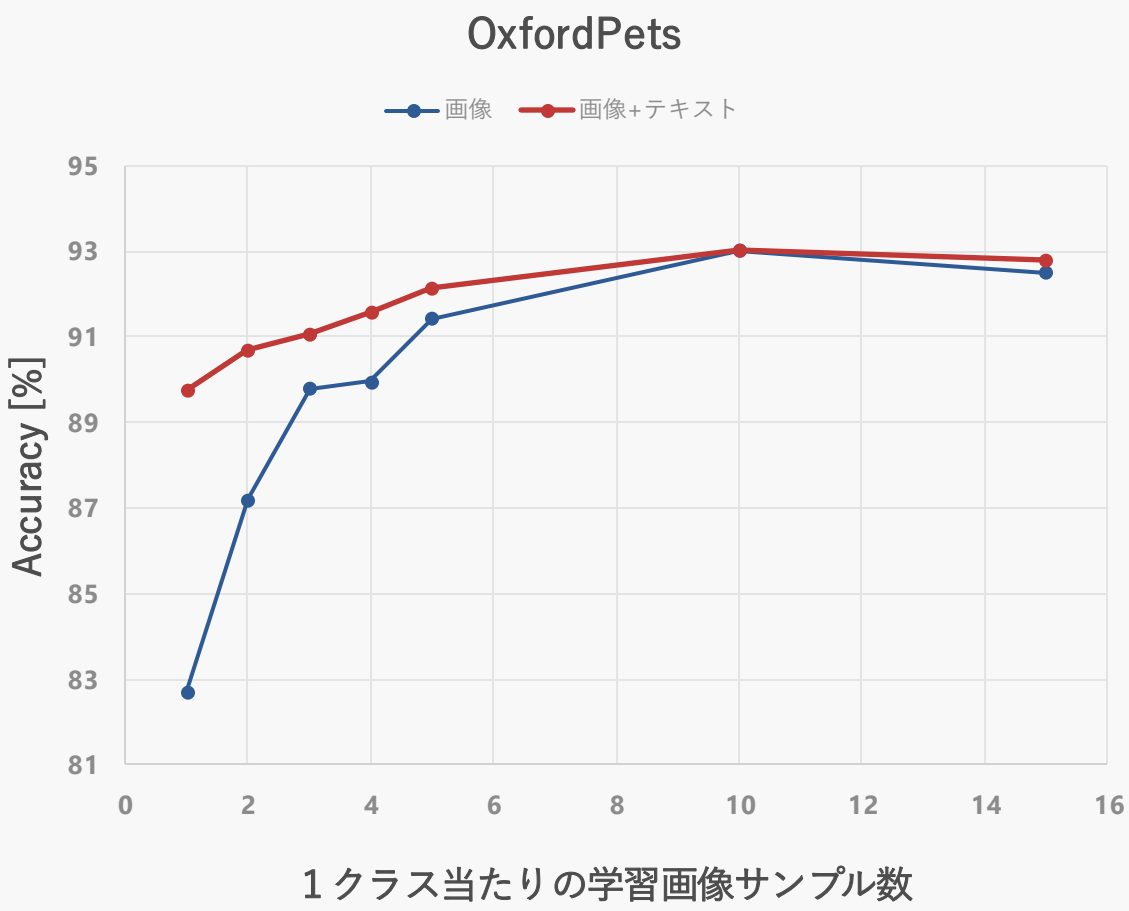
Few-shot Classification

- 学習に用いるクラス毎の画像枚数を1, 2, 3, 4, 5, 10, 15枚として実験
- 加えるテキストの量も変えて実験
  - CUB（1～300個/クラス），OxfordPets（1から1000個/クラス）

実験：Few-shot Classification

- 2つのデータセットでの実験結果





実験：Few-shot Classification

- 2つのデータセットでの実験結果
  - 画像+テキストの列の括弧内は学習に用いたテキストの数

CUB-200-2011

画像枚数	画像	画像+テキスト（個/クラス）
1	68.45	73.41 (t=5)
2	73.88	75.71 (t=2)
3	76.46	76.66 (t=4)
4	77.93	78.02 (t=3)
5	78.78	78.11 (t=1)
10	79.59	79.11 (t=1)
15	79.81	79.71 (t=1)

（$1 \leq t \leq 300$）

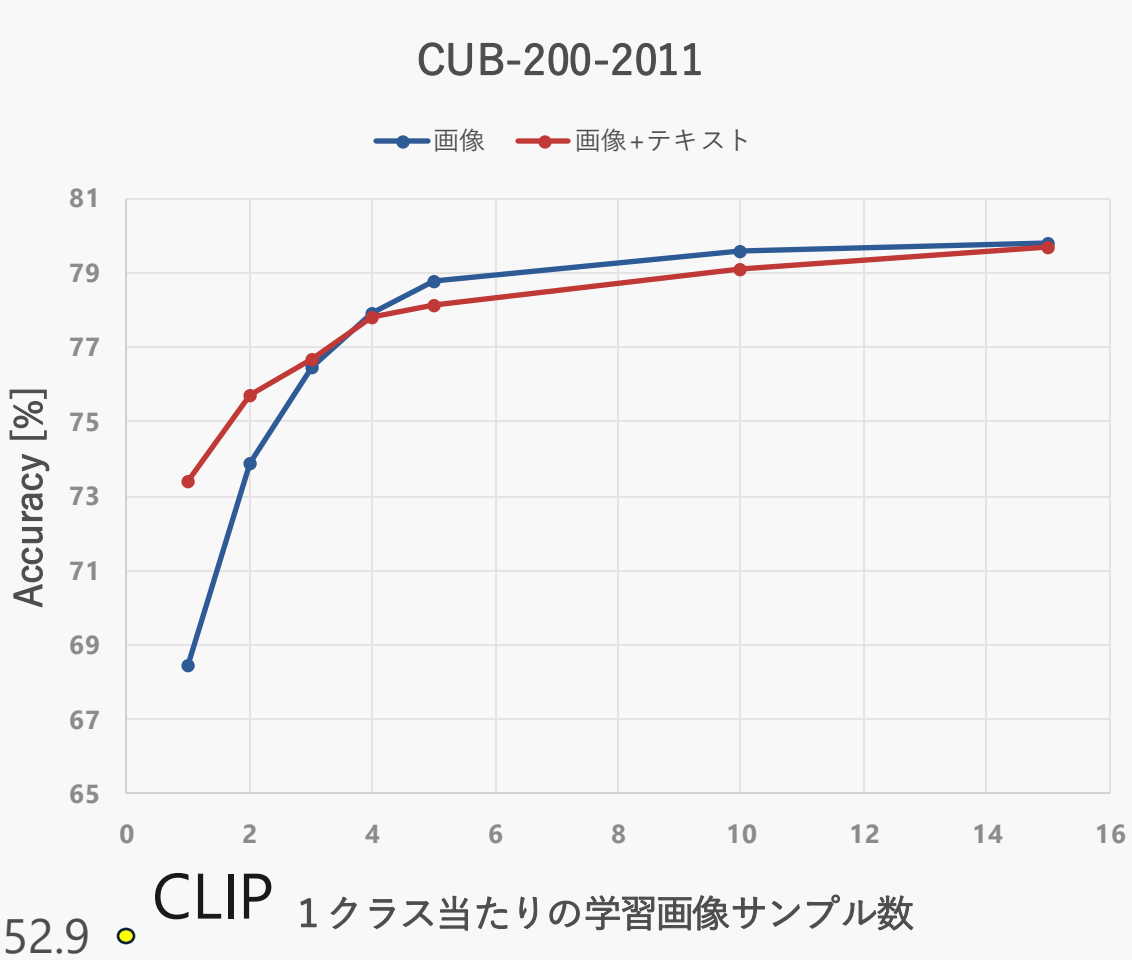
OxfordPets

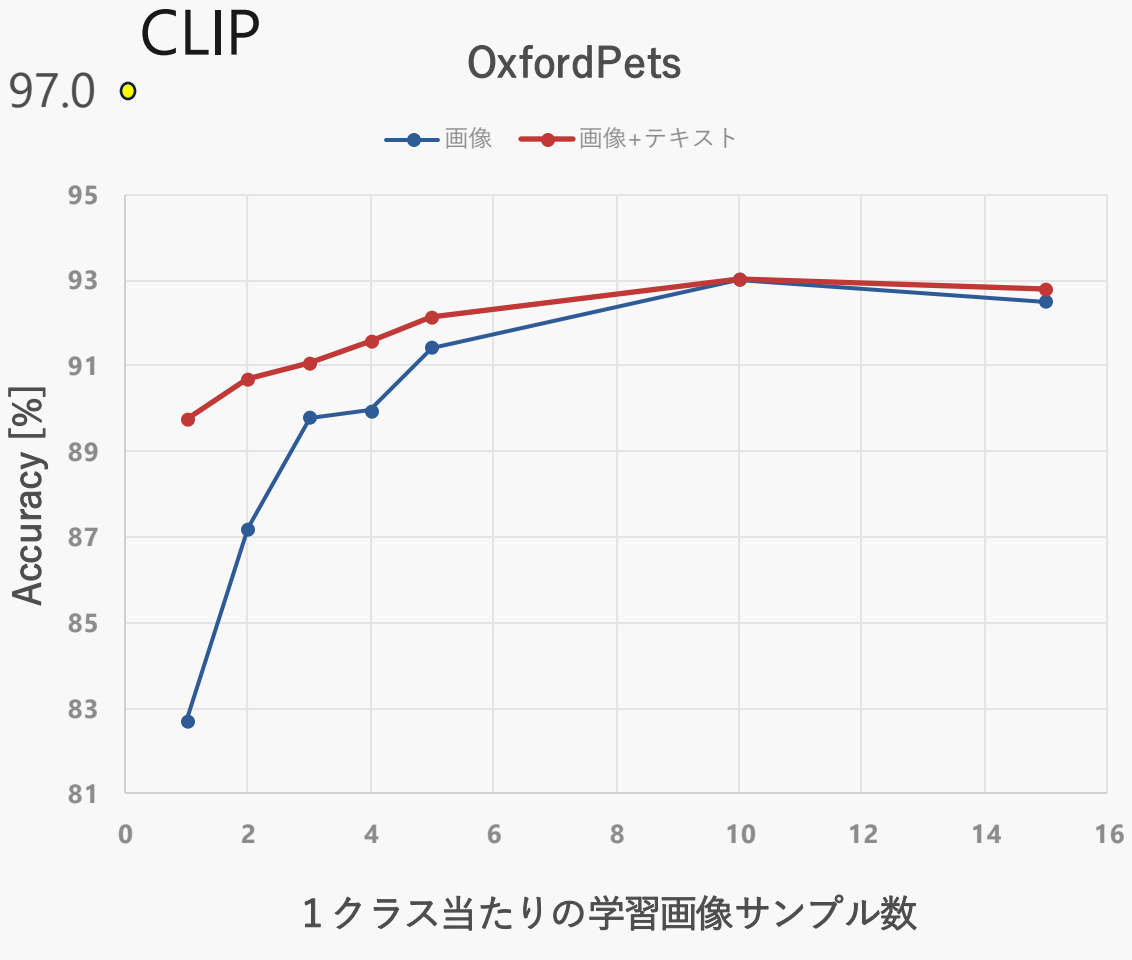
画像枚数	画像	画像+テキスト（個/クラス）
1	82.7	89.78 (t'=20)
2	87.21	90.71 (t'=3)
3	89.8	91.08 (t'=30)
4	89.96	91.59 (t'=5)
5	91.43	92.12 (t'=10)
10	93.01	93.03 (t'=10)
15	92.5	92.82 (t'=30)

（$1 \leq t' \leq 1000$）

実験：Zero-shotとの比較

- 2つのデータセットでの実験結果





目次

まとめ

- LLMを用いてテキストデータセットを作成
- Few-shot設定ではテキストによる学習が有効

今後の予定

- 関連研究の実装と比較

大規模言語モデルを用いた Vison-Language ModelによるFew-shot分類

B4　DL班　川越　壮