10/10 Discussion

論文調査

TR24006 小林 亮太（中部大学工学部ロボット理工学科　藤吉研究室）

http://mprg.jp

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- 再現実験との比較
  - 概要
  - 結果
- 論文調査

実験概要

- MCNを用いて近づけるモーダルの組み合わせ方による学習効果を調査
  - 3モーダルを二段階で学習する実験：V（ビデオ），A（オーディオ），T（テキスト）
    - AV_T：AとVで学習してからTを追加
    - VT_A：VとTで学習してからAを追加
    - AT_V：AとTで学習してからVを追加

- 卒論時は低い精度
  - エポック数が不足
  - 追加の実験が必要

実験条件

- アーキテクチャ　：MCN
- Feature Extractor
  - ビデオ　　：ResNet152
  - オーディオ　：DaveNet
  - テキスト　　：Word2vec
- データセット　　：HowTo100M
  - ビデオ解像度　　：454 × 256
  - ビデオフレームレート　　：30FPS
  - オーディオサンプリングレート　：16kHz
- バッチサイズ　　：128
- エポック数　　：｛再現実験：30｝，　｛組み合わせごとに異なる｝
- 学習率　　：0.0001
- 特徴量次元数　　：4096

実験結果　再現実験との比較

- 追加実験と再現実験での精度の違い

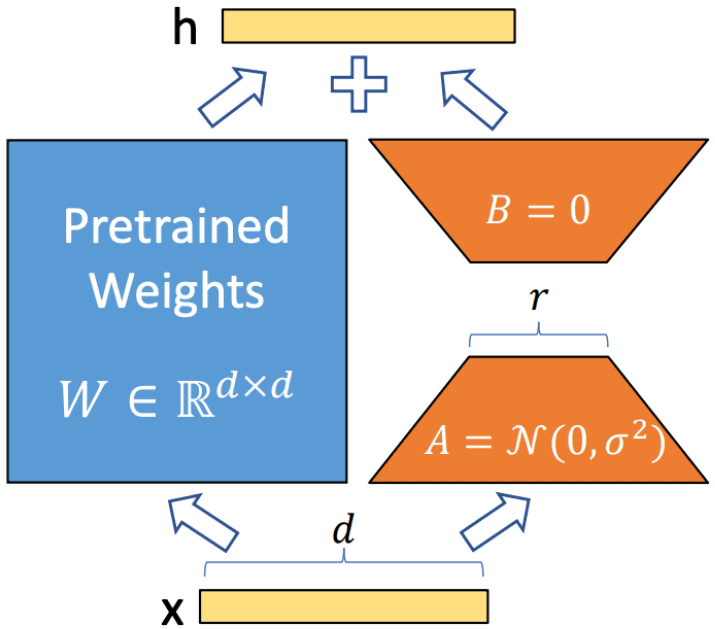
		YouCook			MSR－VTT		
		R@1	R@5	R@10	R@1	R@5	R@10
追加実験	AV_T	1.26	4.18	6.96	2.69	8.06	14.05
	VT_A	6.84	16.2	22.6	6.20	14.5	20.1
	AT_V	8.06	17.3	21.6	1.65	5.27	8.37
再現実験		16.8	31.3	40.2	4.75	12.2	18.4

論文調査

LoRA：Low-Rank Adaptation of Large Language Models[Edward Hu+, ICLR'22]

- 背景
  - 大規模言語モデルの進化
    - パラメータ数の増加
  - ファインチューニングの課題に対処
    - コストの増大
- LoRAの提案
  - 事前学習済みモデルの重みの固定
  - 低ランク分解行列を挿入

$$W + AB$$



LoRA：Low-Rank Adaptation of Large Language Models[Edward Hu+, ICLR'22]

- LoRAの提案

$$W + AB$$

$$\begin{pmatrix} 2 & 3 & 1 & \cdots \\ 6 & 8 & 0 & \cdots \\ 9 & 7 & 4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

(m × n)
固定

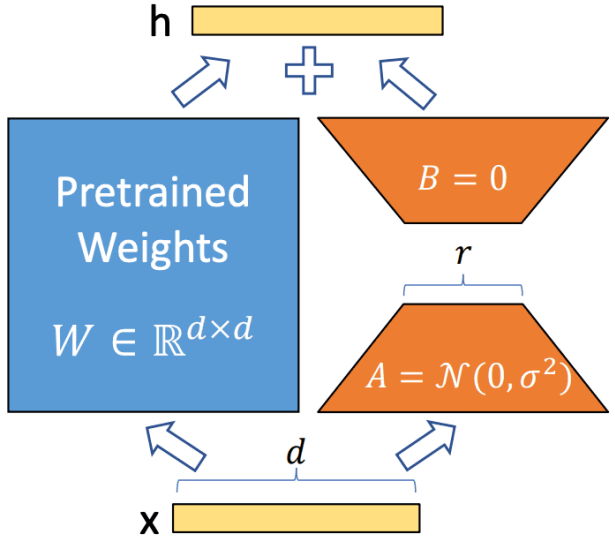
$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}$$

(m × k)

$$(1 \quad 0 \quad 1 \quad 0 \quad \cdots)$$

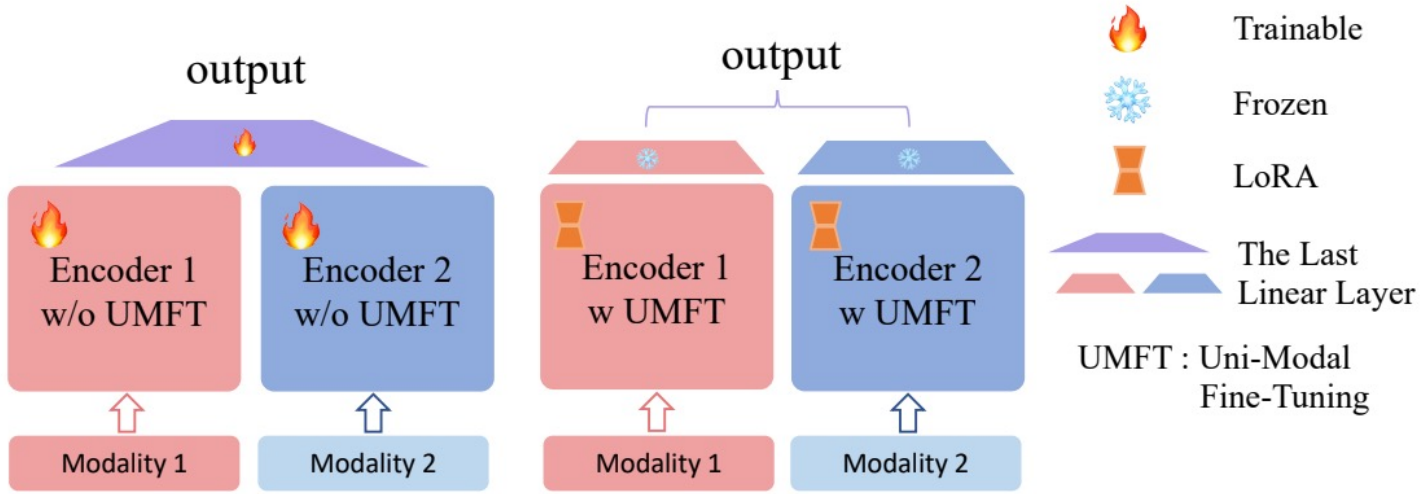
(k × n)

(m × n)



Improving Discriminative Multi-Modal Learning with Large-Scale Pre-Trained Models [C Du, ICLR'24]

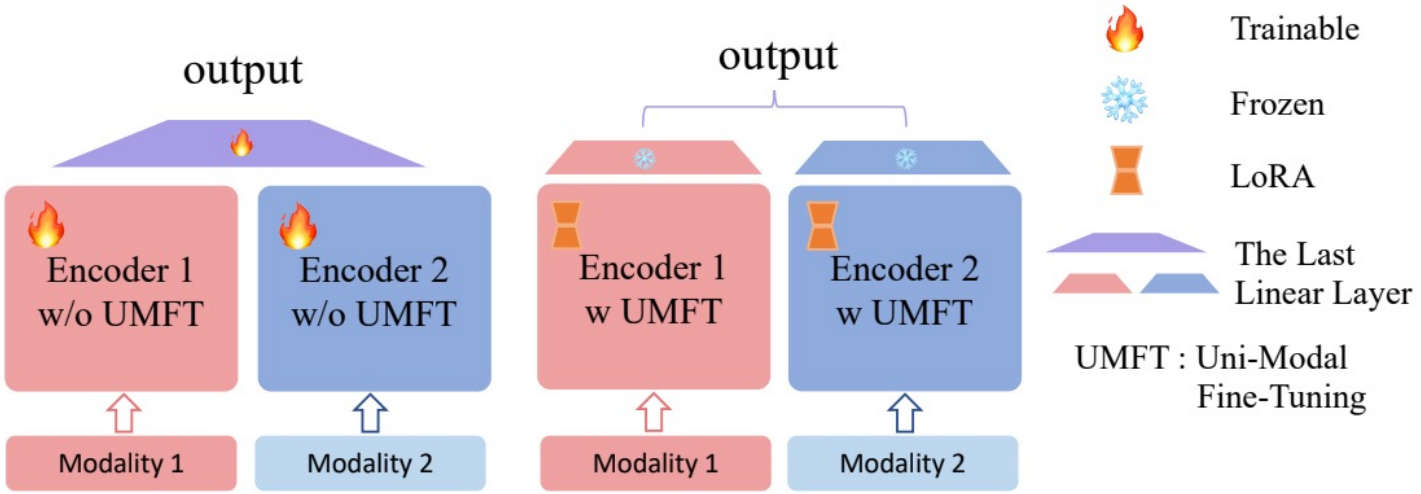
- 各モーダルごとに学習済みの大規模モデルをエンコーダとして使用
- マルチモーダル向けにファインチューニング
- LoRAを適応
  - コストの低減
  - モダリティ競争やモダリティ遅延を抑制



従来　　論文の手法

Improving Discriminative Multi-Modal Learning with Large-Scale Pre-Trained Models [C Du, ICLR'24]

- LoRAを各モーダルのエンコーダに適用
  - 一部または全てのモーダル
- 各エンコーダからの出力の平均から損失を算出



Improving Discriminative Multi-Modal Learning with Large-Scale Pre-Trained Models [C Du, ICLR'24]

- 評価データセット
  - オーディオ，ビジュアル（AVE，Kinetics-Sound，CREMA-D）
  - ビジュアル，テキスト（UPMC Food101，MM-IMDM）
  - 動作（UCF101）

Improving Discriminative Multi-Modal Learning with Large-Scale Pre-Trained Models [C Du, ICLR'24]

- オーディオ，ビジュアル（AVE，Kinetics-Sound，CREMA-D）
  - AVE　　：10秒のビデオクリップ　28カテゴリ
  - Kinetics-Sound　：短いビデオクリップ　32カテゴリ
  - CREMA-D　　：2-3秒のビデオクリップ　6カテゴリ

Top-1 Test Accuracy of different methods on Audio-Visual Datasets

Method	Backbone (A/V)	AVE	KS	CREMA-D
G-Blending (Wang et al., 2020)	ResNet18/ResNet18	65.5	62.2	58.7
OGM-GE (Peng et al., 2022)	ResNet18/ResNet18	76.9	63.1	62.2
PMR (Fan et al., 2023)	ResNet18/ResNet18	74.3	-	65.3
UME* (Du et al., 2023)	ResNet18/ResNet18	85.4	78.8	78.2
MMLoRA (ours)	ResNet18/ResNet18	**86.9**	**79.4**	**81.9**
Multi-Modal Baseline*	ViT-B/ViT-B	94.7	90.6	87.6
Classifier on frozen features*	ViT-B/ViT-B	93.7	90.1	85.3
MBT (Nagrani et al., 2021)	ViT-B/ViT-B	-	85.0	-
UMT* (Du et al., 2023)	ViT-B/ViT-B	93.7	90.3	87.8
OGM-GE* (Peng et al., 2022)	ViT-B/ViT-B	95.5	90.4	88.4
UME* (Du et al., 2023)	ViT-B/ViT-B	95.4	90.8	87.8
Fully Fine-tuned UME*	ViT-B/ViT-B	95.2	91.3	87.5
MMLoRA (ours)	ViT-B/ViT-B	**96.2**	**91.4**	**88.6**

Improving Discriminative Multi-Modal Learning with Large-Scale Pre-Trained Models [C Du, ICLR'24]

- ビジュアル，テキスト（UPMC Food101，MM-IMDM）
  - UPMC Food101　：食品画像とレシピのテキスト　101カテゴリ
  - MM-IMDM　：映画ポスターと映画のプロットの説明テキスト？
- 動作（UCF101）
  - UCF101：RGB画像とOptical-Flow情報　101カテゴリ

UPMC Food101 (Acc.) and MMIMDB(F1-Micro/F1-Macro).

Method	Food101	MM-IMDB
MMBT	92.1	66.8/61.6
MMLoRA (ours)	**93.7**	**67.2/61.7**
Baseline	93.29	64.9/59.6
PMF	91.51	64.5/58.8
PMF-L	91.68	66.7/61.7
MBT	93.6	64.8/59.6
MMBT	94.10	66.1/60.8
MMLoRA (ours)	**95.9**	**71.7/67.5**

Top-1 Test Accuracy (in %) of different methods on UCF101.

Method	Backbone	Acc.
MM Baseline	Res18/Res18	82.3
G-Blending	Res18/Res18	83.0
OGM-GE	Res18/Res18	84.0
UMT	Res18/Res18	84.5
UME	Res18/Res18	86.8
MMLoRA	Res18/Res18	**87.1**
UME*	ViT-B/Res18	93.0
MMLoRA	ViT-B/Res18	**93.4**

BRAVE : Broadening the visual encoding of vision-language models [O Kar+, ECCV'24]

HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts [W Kim+, ECCV'24]

おわりに

- 追加実験結果　　：MTGで相談

- 今後の予定：
  - 論文調査
  - 今後の方針を固める

資料

Improving Discriminative Multi-Modal Learning with Large-Scale Pre-Trained Models [C Du, ICLR'24]

- ビジュアル，テキスト（UPMC Food101，MM-IMDM）
  - UPMC Food101　：食品画像とレシピのテキスト　101カテゴリ
  - MM-IMDM　：映画ポスターと映画のプロットの説明テキスト？
    - {全サンプルに対する平均的な性能} / {各クラスごとの性能の平均}
- 動作（UCF101）
  - UCF101：RGB画像とOptical-Flow情報　101カテゴリ

UPMC Food101 (Acc.) and MMIMDB(F1-Micro/F1-Macro).

Method	Food101	MM-IMDB
MMBT	92.1	66.8/61.6
MMLoRA (ours)	**93.7**	**67.2/61.7**
Baseline	93.29	64.9/59.6
PMF	91.51	64.5/58.8
PMF-L	91.68	66.7/61.7
MBT	93.6	64.8/59.6
MMBT	94.10	66.1/60.8
MMLoRA (ours)	**95.9**	**71.7/67.5**

Top-1 Test Accuracy (in %) of different methods on UCF101.

Method	Backbone	Acc.
MM Baseline	Res18/Res18	82.3
G-Blending	Res18/Res18	83.0
OGM-GE	Res18/Res18	84.0
UMT	Res18/Res18	84.5
UME	Res18/Res18	86.8
MMLoRA	Res18/Res18	**87.1**
UME*	ViT-B/Res18	93.0
MMLoRA	ViT-B/Res18	**93.4**

現在の実験後の予定

- 未定

- 岡本先輩からのコメント
  - 対照学習時にLoRAのような形でくっつけて「LoRAを介して特定のモーダル間を近づけることに特化した学習」をモーダルの組み合わせごとに同時に行って，共通の特徴量の学習と下流タスクに応じてLoRAを使い分けることで各下流タスクで高い精度を発揮可能な仕組みみたいなこと
- 継続学習を勧めらている
- 最適輸送