10/17 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

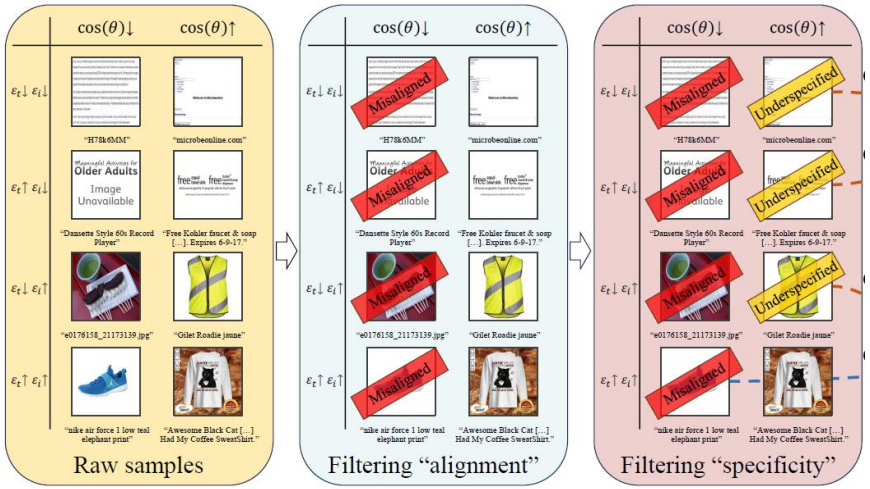
はじめに

- 論文調査
  - HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts
  - ~~TOWARDS META-PRUNING VIA OPTIMAL TRANSPORT~~

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

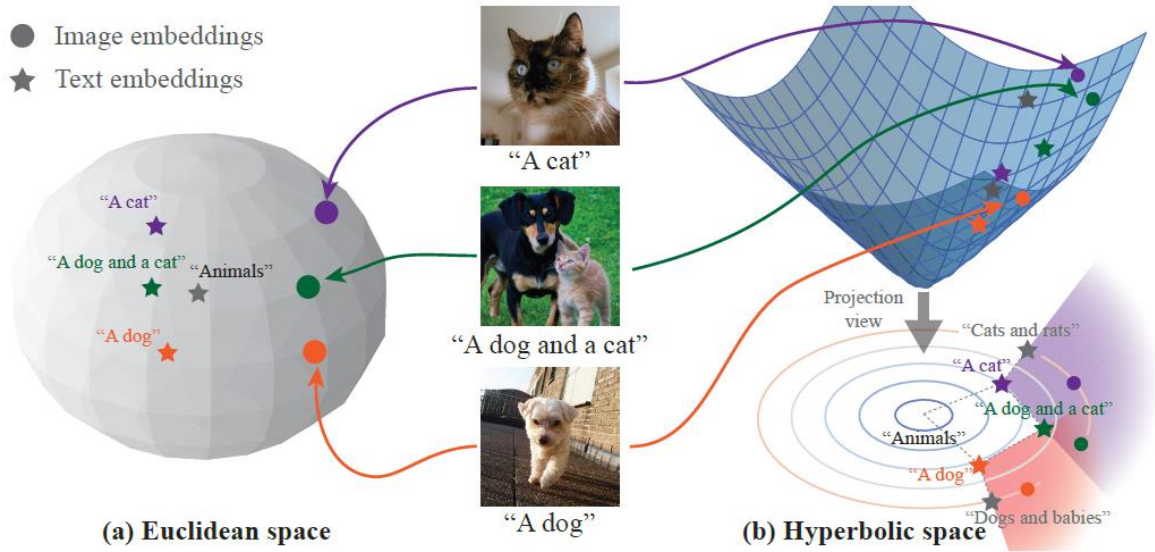
- 自己教師あり学習ではラベル付けされていないデータを使用
- 大量のデータが必要
  - 人間によってラベル付け・検証された高品質なデータセットと同等になるまでの量
  - 計算コストとストレージサイズに問題が発生

- HYPE：双曲線空間を用いたデータセットのフィルタリング



双曲線空間

- 双曲線空間：負の曲率を持つ「曲がった」空間
  - 中心から離れるほど指数関数的に広がる
- MERU：双曲線空間を使うマルチモーダル向けCLIP
  - 画像とテキストを双曲線空間へ埋め込む



ユークリッド空間と双曲線空間の比較

HYPEの仕組み

- HYPEスコア
  - スコアに基づいて，データサンプルをランク付け
  - 上位のサンプルを選択してフィルタリング済みデータセットを作成

$$HYPE_{score} = \varepsilon_i + \varepsilon_t - d_L + cos\theta$$

- ε_i，ε_t　：画像，テキストの特異性
- d_L　：ローレンツ距離
- $cos\theta$　：コサイン類似度

フィルタリング：特異性ε_i ε_t

- 画像もしくはテキストの特異性
- ε_i：画像特異性
  - 各画像が参照テキストをどの程度含んでいるか
- ε_t：テキスト特異性
  - 各テキストが参照画像をどの程度含んでいるか

- 双曲線空間内で各データ点を中心とする「円錐」の範囲
  - 多くのデータを含んでいるほど低い特異性

フィルタリング：ローレンツ距離d_L ・コサイン類似度$cos\theta$

- 画像とテキストのペアが十分に類似しているかをチェック

- d_L　：ローレンツ距離
  - 双曲線空間内において 2 点間の距離を測る方法
  - 負の値 $(-d_L)$ を使用：値が大きいほど近いことを示す
  - 中心から離れるほど距離の増加が加速

- $cos\theta$ ：コサイン類似度
  - CLIPをそのまま使用
  - ユークリッド空間内において 2 点間の距離を測る方法

評価

- DataCompベンチマークで評価
  - DataCompデータセットを使用
    - small(12.8M)，medium(128M)，large(1.28B)，xlarge(12.8B)の４つのスケール
  - モデルや学習方法を固定
    - フィルタリングされたデータセットの品質を評価可能
  - 38のタスクが存在：ImageNet分類，VTAB，検索タスク，etc…
  - 全タスクでのゼロショットのスコアの平均で比較

- smallとmediumを使用

評価：結果

Method	Datacomp Scale	Sample Size	Uniform	ImageNet	ImageNet Dist. Shift	VTAB	Retrieval	Average
CLIP L/14 30% [20]	Small	3.8M	Yes	0.051	0.055	0.190	0.119	0.173
WS [31]	Small	4.1M	Yes	0.056	0.061	0.196	0.132	0.180
HYPE 10%	Small	1.2M	Yes	0.051	0.056	0.162	0.102	0.150
HYPE 20%	Small	2.3M	Yes	0.064	0.063	0.190	0.130	0.176
HYPE 30%	Small	3.5M	Yes	0.054	0.057	0.182	0.133	0.170
CLIP L/14 30% [20]	Medium	38.0M	Yes	0.273	0.230	0.338	0.251	0.328
WS [31]	Medium	24.8M	Yes	0.305	0.253	0.363	0.270	0.342
T-MARS [45]	Medium	23.0M	No	0.338	0.274	0.371	0.231	0.357
CIDS [76] *	Medium	21.3M	No	0.326	0.262	0.372	0.258	0.365
DFN [19] *	Medium	17.1M	Yes	<u>0.376</u>	<u>0.300</u>	0.384	0.284	0.372
HYPE 10%	Medium	11.6M	Yes	0.327	0.257	0.365	0.246	0.340
HYPE 20%	Medium	23.1M	Yes	0.338	0.269	0.357	<u>0.286</u>	0.343
HYPE 30%	Medium	34.7M	Yes	0.300	0.243	0.337	0.276	0.332
HYPE 10% + CIDS [76] *	Medium	18.9M	No	0.346	0.276	<u>0.390</u>	0.264	<u>0.373</u>
HYPE 10% + DFN [19] *	Medium	21.5M	No	**0.382**	**0.303**	**0.393**	**0.306**	**0.379**

評価：比較手法

- CLIP L/14 ：使用するモデル
- 30%　　：データセットからランダムに30%選択
- WS　　：複数のフィルタリング基準を組み合わせて使用
  - 言語フィルタ　　：英語のテキストのみ選択
  - テキスト長フィルタ　：一定以上の長さのテキストを選択
  - 画像サイズフィルタ　：一定以上のサイズの画像を選択
  - コサイン類似度　　：画像とテキストの対応度が高いものを選択
- T-MARS　　：テキスト領域をマスクした画像を使用
  - マスク画像とオリジナル画像それぞれのテキストとのコサイン類似度の比較
    - スコアの変化が小さいほど高品質と定義
- DFN　　：高品質なデータを識別するように訓練されたモデルを使用

評価：結果

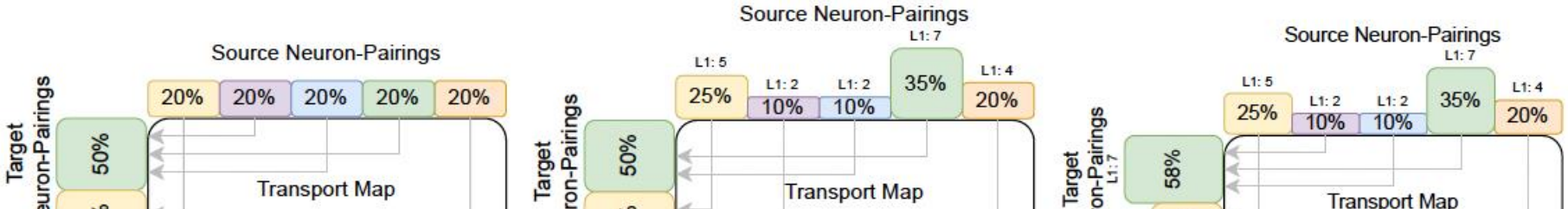
- HYPE 10% + DFNが最高
  - データセット全体にHYPEスコアを適用 → スコアの上位１０％のサンプルを選択
  - 残りの９０％に対してDFNを適用 → 追加のサンプルを選択

Method	Datacomp Scale	Sample Size	Uniform	ImageNet	ImageNet Dist. Shift	VTAB	Retrieval	Average
CLIP L/14 30% [20]	Small	3.8M	Yes	0.051	0.055	0.190	0.119	0.173
WS [31]	Small	4.1M	Yes	0.056	0.061	0.196	0.132	0.180
HYPE 10%	Small	1.2M	Yes	0.051	0.056	0.162	0.102	0.150
HYPE 20%	Small	2.3M	Yes	0.064	0.063	0.190	0.130	0.176
HYPE 30%	Small	3.5M	Yes	0.054	0.057	0.182	0.133	0.170
CLIP L/14 30% [20]	Medium	38.0M	Yes	0.273	0.230	0.338	0.251	0.328
WS [31]	Medium	24.8M	Yes	0.305	0.253	0.363	0.270	0.342
T-MARS [45]	Medium	23.0M	No	0.338	0.274	0.371	0.231	0.357
CIDS [76] *	Medium	21.3M	No	0.326	0.262	0.372	0.258	0.365
DFN [19] *	Medium	17.1M	Yes	<u>0.376</u>	<u>0.300</u>	0.384	0.284	0.372
HYPE 10%	Medium	11.6M	Yes	0.327	0.257	0.365	0.246	0.340
HYPE 20%	Medium	23.1M	Yes	0.338	0.269	0.357	<u>0.286</u>	0.343
HYPE 30%	Medium	34.7M	Yes	0.300	0.243	0.337	0.276	0.332
HYPE 10% + CIDS [76] *	Medium	18.9M	No	0.346	0.276	<u>0.390</u>	0.264	<u>0.373</u>
HYPE 10% + DFN [19] *	Medium	21.5M	No	**0.382**	**0.303**	**0.393**	**0.306**	**0.379**

TOWARDS META-PRUNING VIA OPTIMAL TRANSPORT [A.Theus+,ICLR'24]

- 枝刈り：重要ではないパラメータを削除
  - モデルのサイズを縮小
  - 単純な削除では精度低下を起こすケースが存在

- 提案手法：Intra-Fusion
  - パラメータの削除ではなく融合
  - 融合に最適輸送を活用　→　情報の損失を抑制



Intra-Fusion

- 依存関係のあるパラメータのグループを特定
- 各グループごとに各パラメータの重要度を計算

おわりに

- 論文調査　：HYPE 大規模なデータセットのフィルタリング

- 今後の予定 ：
  - 論文調査
    - Low-Rank Rescaled Vision Transformer Fine-Tuning: A Residual Design Approach
  - 調査した手法のコードの有無確認・再現実験
  - 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

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