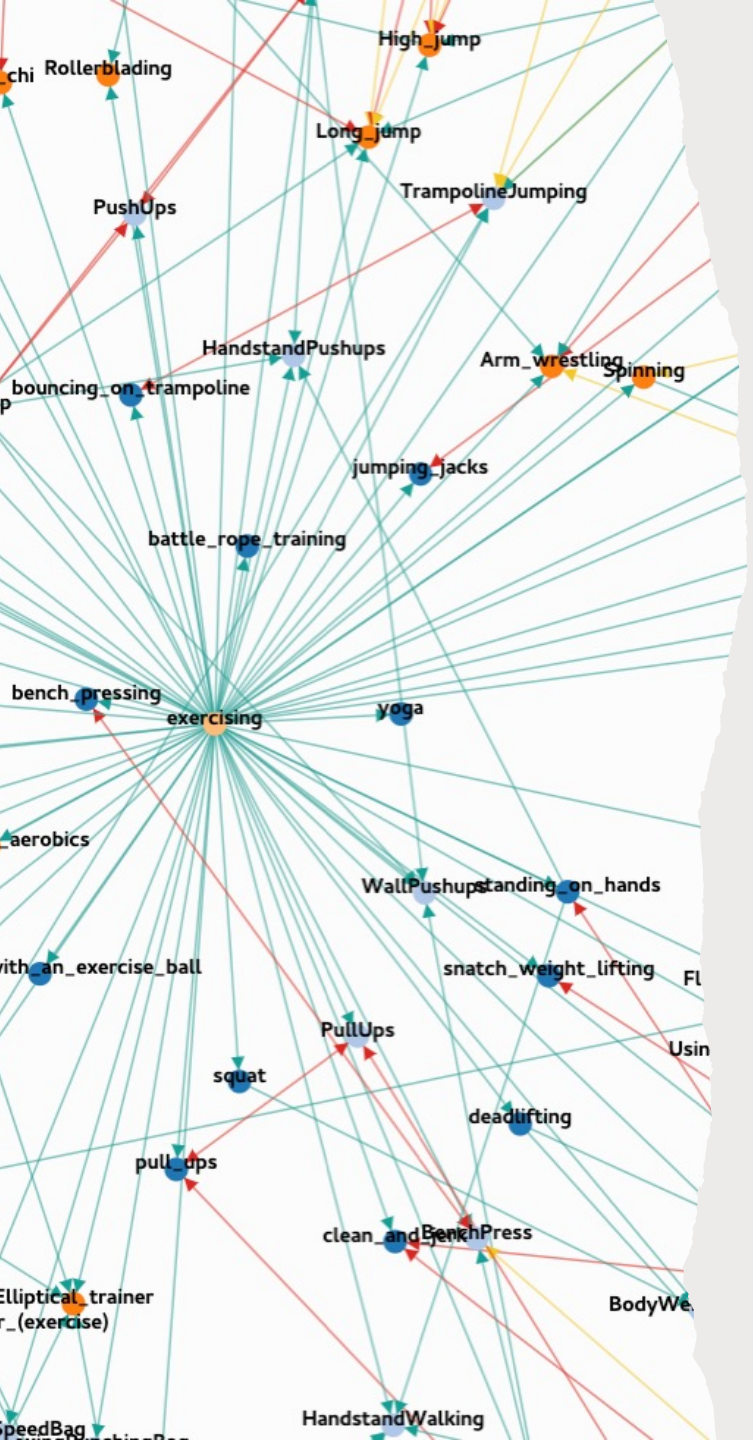


産総研AIセミナー

メタ動画データセットによる動作認識の現状と可能性

千葉工業大学
人工知能・ソフトウェア技術研究センター
吉川 友也

https://yuya-y.com

自己紹介

吉川 友也 (よしかわ ゆうや)

千葉工業大学
人工知能・ソフトウェア技術研究センター
上席研究員

博士（工学）

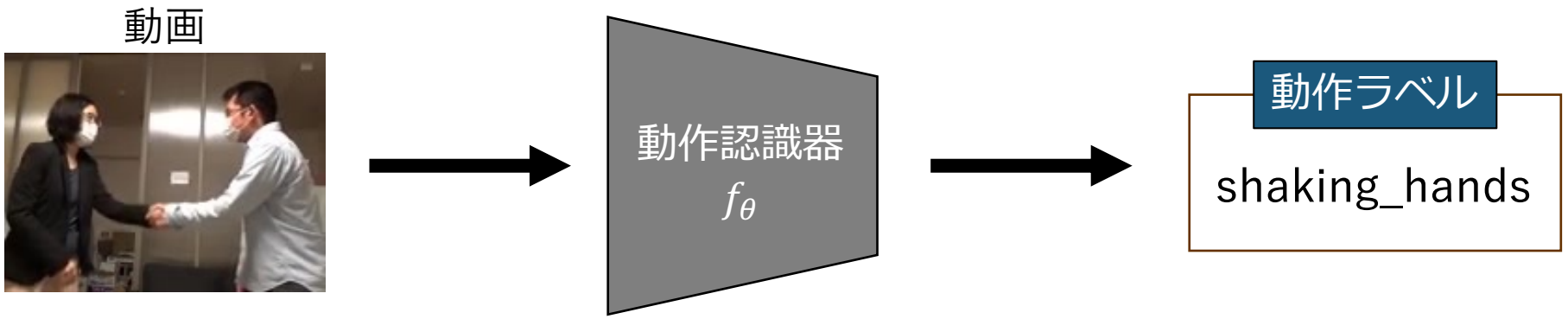
- 2015年 奈良先端科学技術大学院大学

最近の研究トピック

- 説明可能AI (解釈可能な機械学習)
- **動作認識**

人物動作認識（Human Action Recognition）

動作認識器f_θを用いて動画中の人物の動作を分類



研究の目的　　高精度の動作認識器f_θを学習

本講演における研究方針

学習に用いるデータセットを工夫して動作認識器の精度を改善することを目指す

動作認識モデルを学習するためのデータセットの例

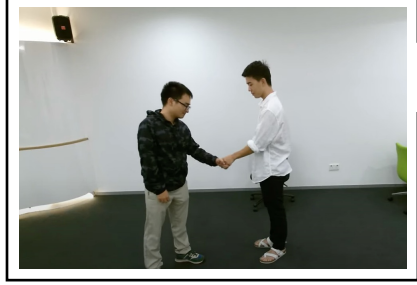
- 大規模・一般動作データセット（事前学習にも使われる）
  - Kinetics-700
  - ActivityNet
- ベンチマークデータセット（手頃なサイズ）
  - UCF101
  - HMDB51
- ドメイン特化
  - EPIC-KITCHENS：料理作業動作
  - Sports-1M：スポーツの種類

	データ数	クラス数
Kinetics-700	530K	700
ActivityNet	21K	200
UCF101	13K	101
HMDB51	5K	51
EPIC-KITCHENS	90K (action segments)	97 verbs 300 nouns
Sports-1M	1133K	487

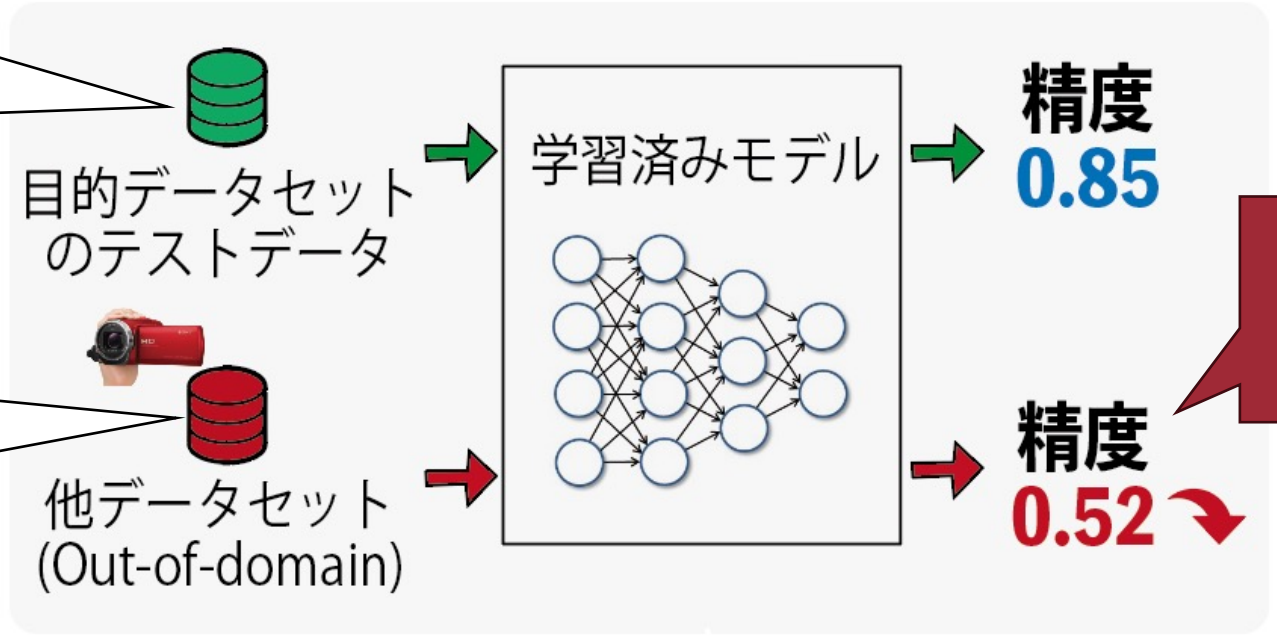
単一のデータセットの限界

- 多くの論文では単一のデータセット内で訓練・テストを行い
  モデルの性能を評価
- しかし、動画のドメインが変わったときに認識できるのかが不明

STAIR Actionsより



NTU RGB+Dより



独自データセットでモデルを学習したい場合

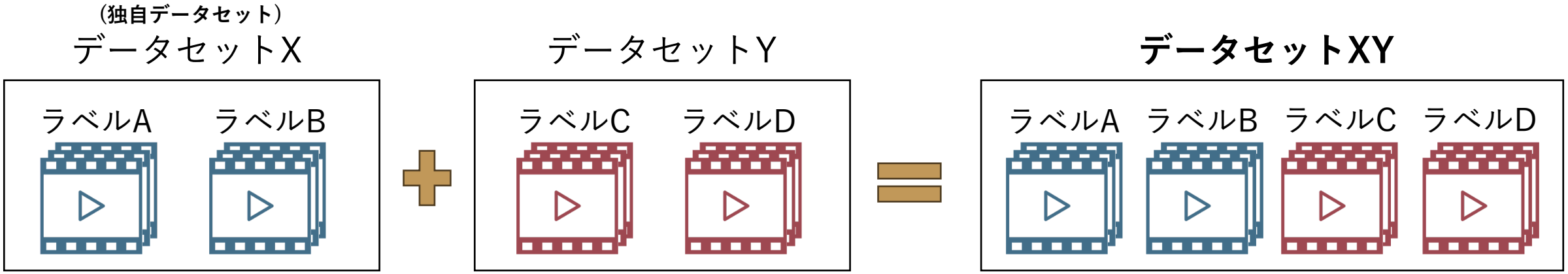
- 独自の動作を認識できるようにしたい場合、自ら動画収集・ラベル付けを行う必要がある

難しいポイント

- データ作成コストが高く
  大規模な教師データを作れない
- ドメインが限定的になりやすい
  - 例：動画撮影方法が限られる、
    登場人物が限られる、等

→ **学習不足（underfitting）で汎化しない可能性**

複数データセットで学習すれば解決？



メリット

- 多様なドメイン、大量のデータで学習することで良い表現学習ができる

デメリット

- モデル全体を更新して表現学習までしようとすると訓練コストが高い
- 各ラベルのデータは増えていないので、モデルの最終層のみの学習では精度向上に繋がらない可能性

課題のまとめ

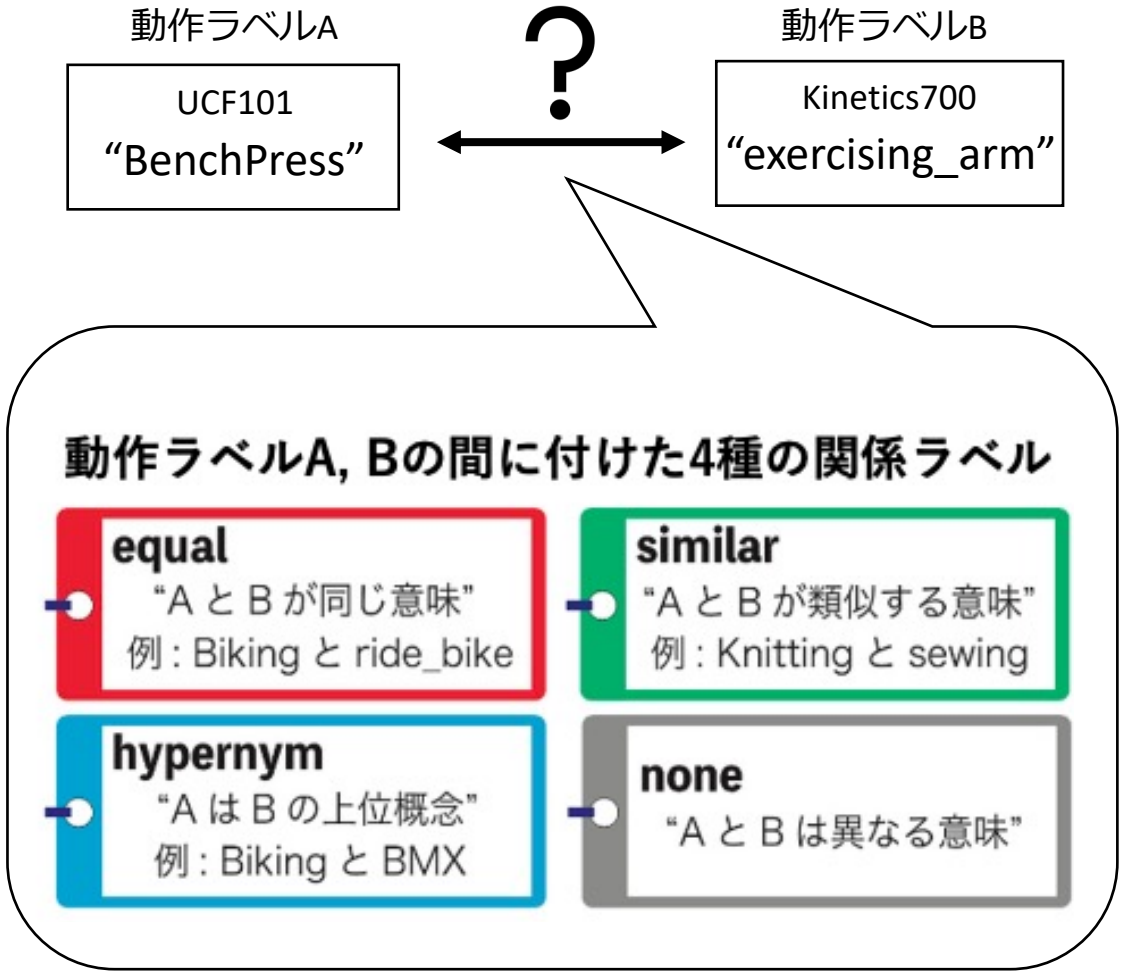
- 単一データセット
  - 他のデータセットの動画に対する認識性能がわからない
- 独自データセット
  - データ作成コストが高く、大規模データが作れない
  - ドメインが限定的になりやすい
- 複数データセット
  - 訓練コストが高い
  - モデルの最終層のみの学習では精度向上に繋がらない可能性

MetaVD (Meta Video Dataset) [Yoshikawa+ 2021]

既存の動作認識データセットの間で動作ラベルの関係性を人手アノテーション

- 6種類のデータセットから構成
  - UCF101 (101クラス)
  - HMDB51 (51クラス)
  - ActivityNet (200クラス)
  - STAIR Actions (100クラス)
  - Charades (157クラス)
  - Kinetics-700 (700クラス)

  今後、更に追加予定



MetaVD (Meta Video Dataset) [Yoshikawa+ 2021]

既存の動作認識データセットの間で動作ラベルの関係性を人手アノテーション

動作ラベルA

UCF101
“BenchPress”

?

動作ラベルB

Kinetics700
“exercising_arm”

全ての動作ラベルペアにアノテーション

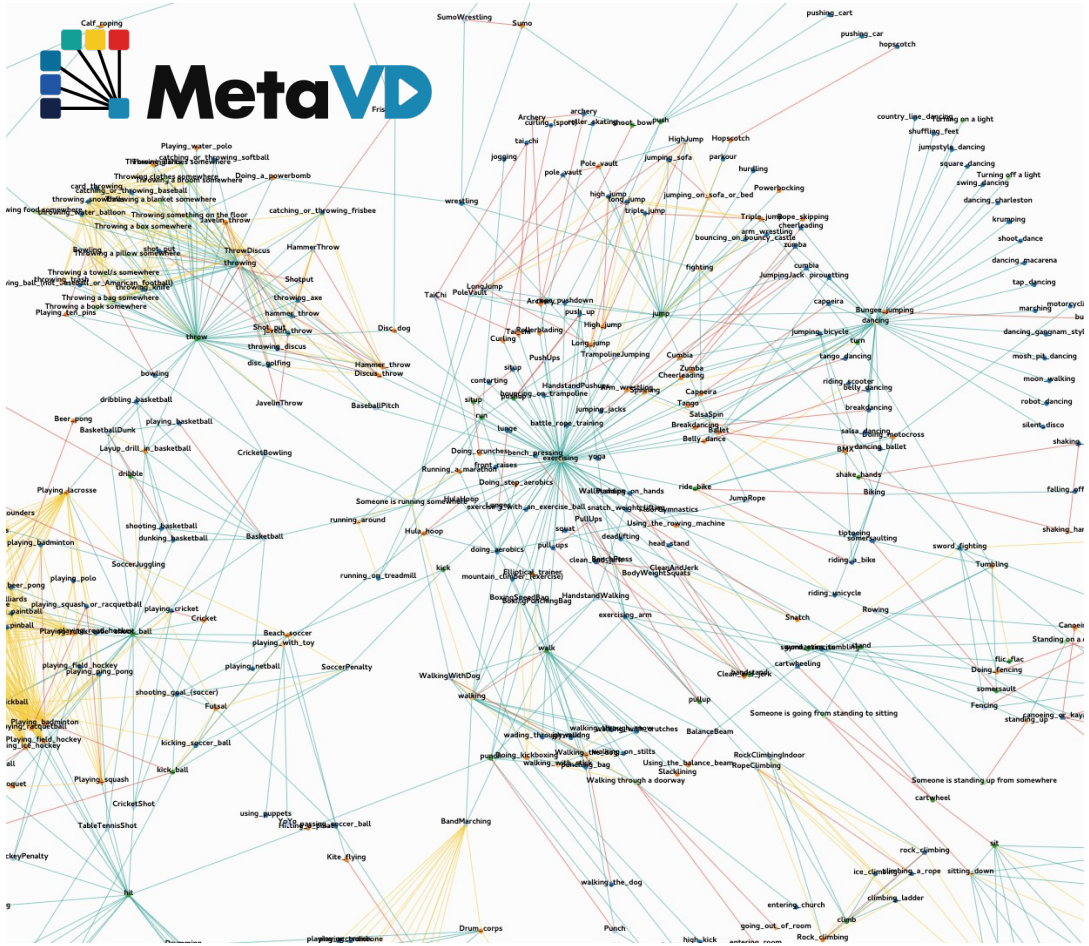
動作ラベルA, Bの間に付けた4種の関係ラベル

equal
“A と B が同じ意味”
例 : Biking と ride_bike

similar
“A と B が類似する意味”
例 : Knitting と sewing

hypernym
“A は B の上位概念”
例 : Biking と BMX

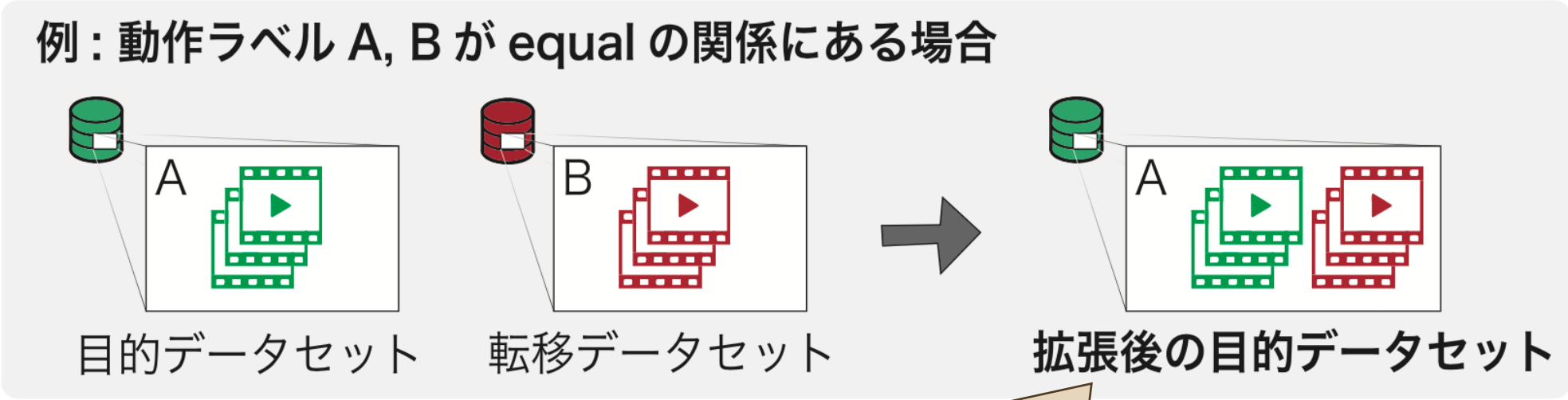
none
“A と B は異なる意味”



https://metavd.stair.center/visualizer.html

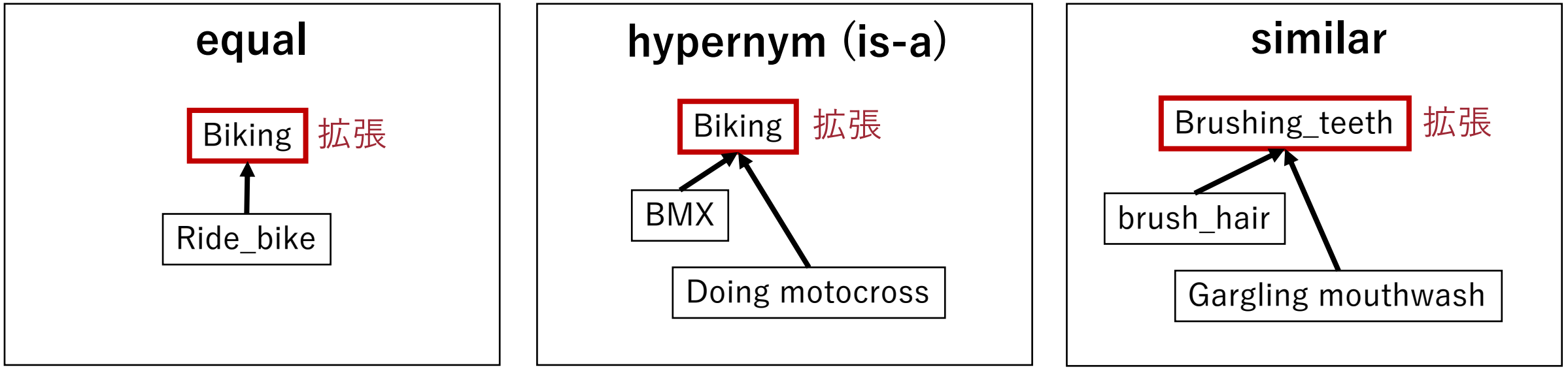
MetaVDを用いたデータセット拡張

MetaVD内の1つのデータセット（目的DS）をその他のデータセット（転移DS）で拡張



転移DSの動作ラベルBの動画を
目的DSの動作ラベルAの動画として追加

関係の種類による拡張される動画の違い



動画の増え方

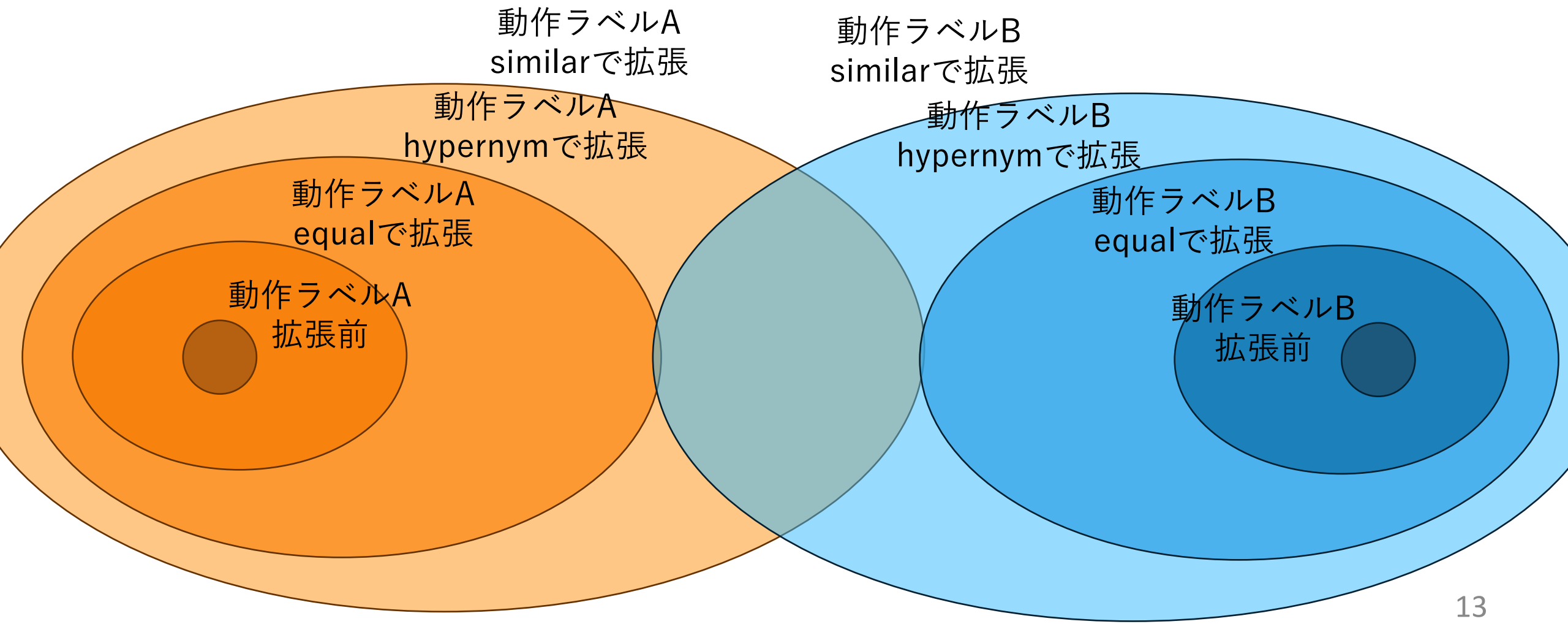
小 → 大

動画のバリエーション

低 → 高

関係の種類による拡張される動画の違い

動作ラベルA, Bが拡張されたときの入力空間の分布のイメージ



MetaVDでデータセット拡張後のサイズ

表. MetaVDにより拡張した各データセットの訓練データ数

Target dataset	equal	equal + similar	equal + is-a	All
UCF101	66,719 (59)	380,560 (77)	107,391 (68)	421,232 (83)
HMDB51	40,996 (30)	68,398 (38)	167,803 (48)	195,205 (51)
ActivityNet	130,315 (121)	720,711 (159)	187,327 (143)	777,723 (168)
STAIR Actions	151,412 (56)	279,994 (76)	327,171 (76)	455,753 (86)
Charades	98,763 (28)	197,025 (67)	160,045 (38)	258,307 (71)
Kinetics-700	587,272 (162)	858,396 (286)	614,708 (196)	885,832 (312)

※ ()内の数字は拡張された動作クラスの数

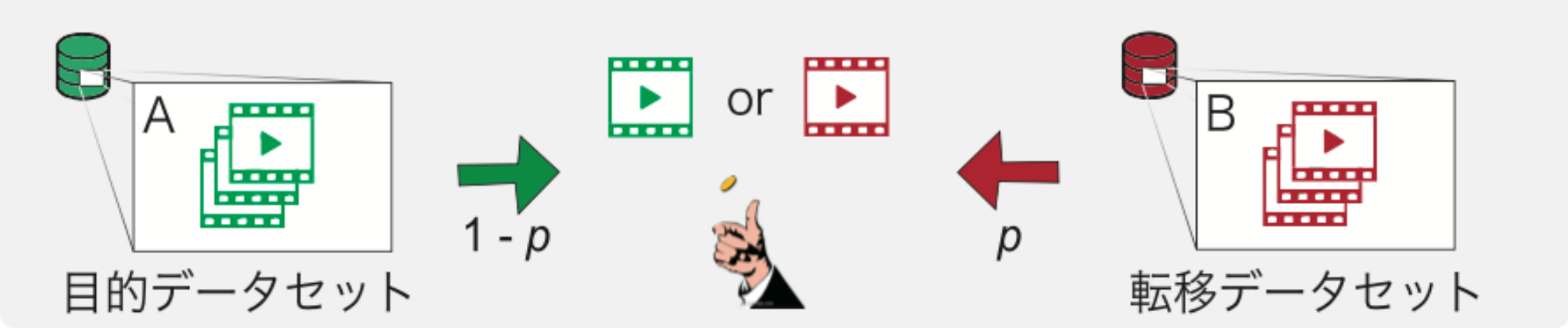
注意：すべての動作ラベルで動画が増えるわけではない

➡ クラス分布が変わり、精度を下げる要因になる

訓練時のミニバッチ構築の戦略

クラス分布を変えず、転移データセットの割合を調整できるようにするため
通常通り目的データセットでミニバッチを作った後、確率pで関係する動作ラベルが付いた転移データセットの動画に置き換える

例：動作ラベルA,Bがequalの関係にあるとき



課題に対するMetaVDを用いた解決策

- 単一データセット

他のデータセットにある関係する動作ラベルの動画を用いて認識性能を評価可能

- 独自データセット

  他のデータセットから、多様なドメインの動画を多数取り込むことが可能

- 複数データセット

目的データセットに必要な動画のみを取り込むことで訓練コストを抑え
各ラベルの動画を増やして精度向上に繋げる

認識性能

$p = 0.5$の場合

UCF101のテスト精度

UCF101のみ	拡張 (equal)	拡張 (equal + is-a)	拡張 (equal + similar)
91.38	91.14	90.99	90.46

UCF101のみとほぼ変わらず

拡張UCF101のテスト精度

	テストデータ		
訓練データ	拡張 (equal)	拡張 (equal + is-a)	拡張 (equal + similar)
UCF101のみ	72.89	63.11	28.66
拡張UCF101	82.43	76.69	33.92

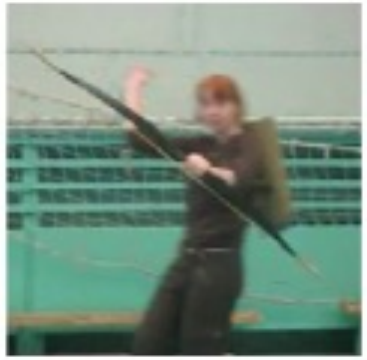
UCF101のみよりも大幅に精度が向上
多様な動画に対する認識性能が向上していることが示唆される

認識結果の例　～ UCF101のみで学習した場合 ～

bowling
(Kinetics-700)

fencing
(HMDB51)

Archery
(UCF101)



UCF101のみで学習

動作認識器
f_θ

UCF101でのテスト精度
91.38%

拡張UCF101でのテスト精度
72.89%

MoppingFloor ✗

SalsaSpin ✗

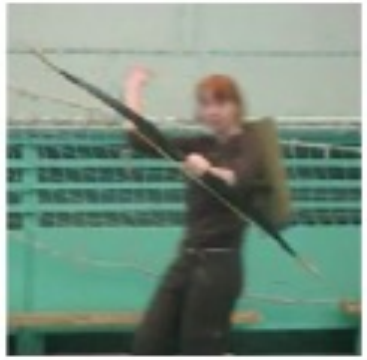
PlayingViolin ✗

認識結果の例　～ 拡張したUCF101で学習した場合 ～

bowling
(Kinetics-700)

fencing
(HMDB51)

Archery
(UCF101)



equal関係で拡張した
UCF101で学習

動作認識器
f_θ

Bowling

Fencing

Archery

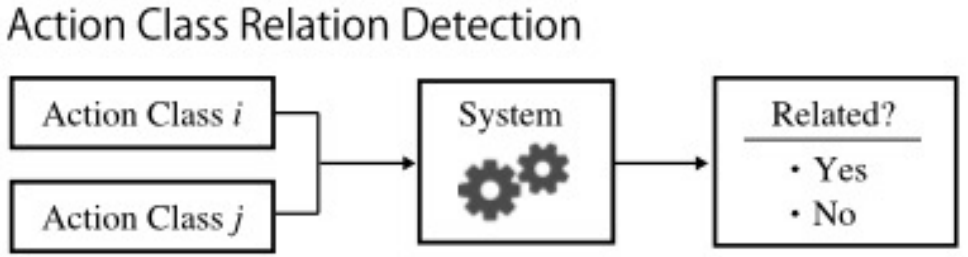
UCF101でのテスト精度
91.14% →

拡張UCF101でのテスト精度
82.43% ↑

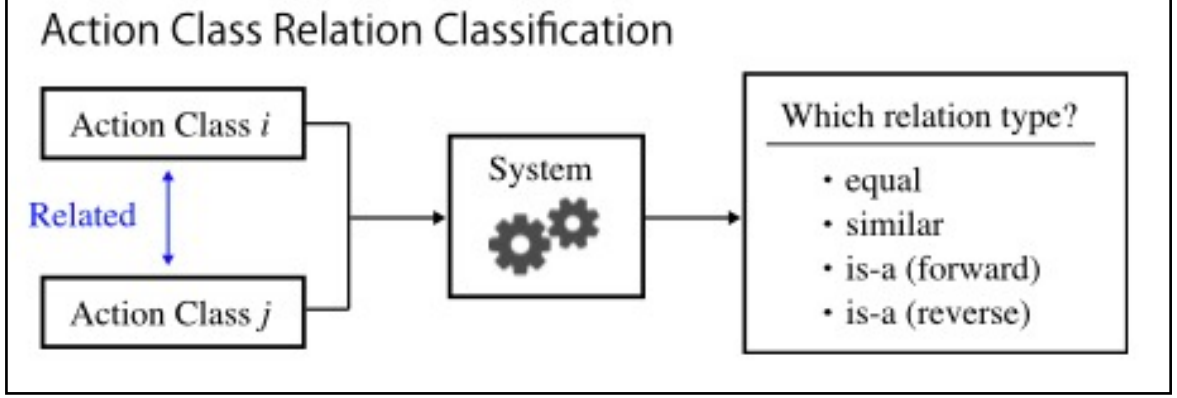
動作ラベルの関係予測 [Yoshikawa+ 2023]

2つの動作ラベルの間にどのような関係があるかを予測

関係の有無を予測するタスク



関係の種類を予測するタスク

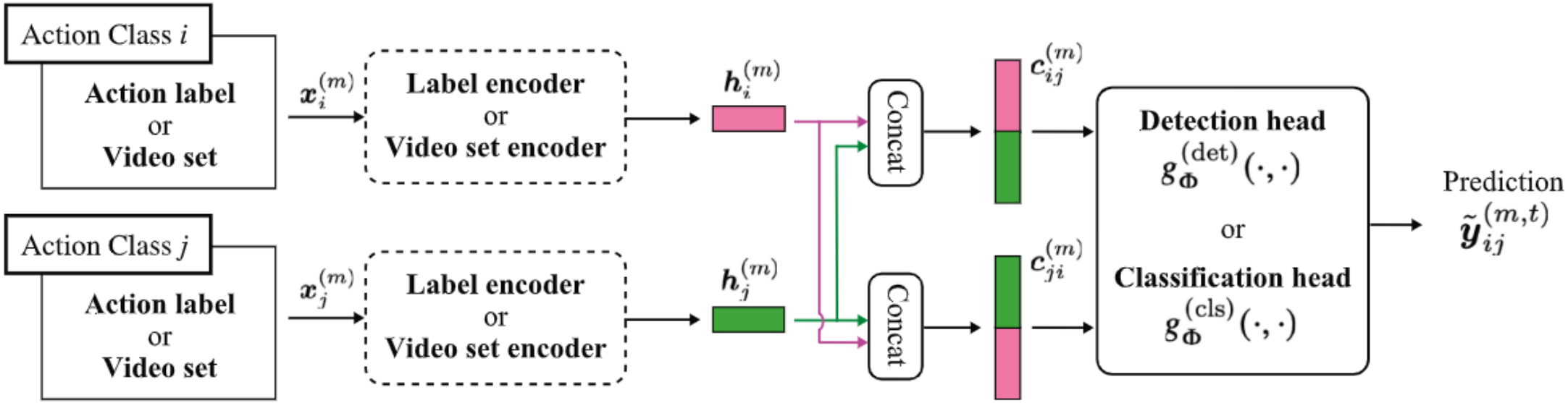


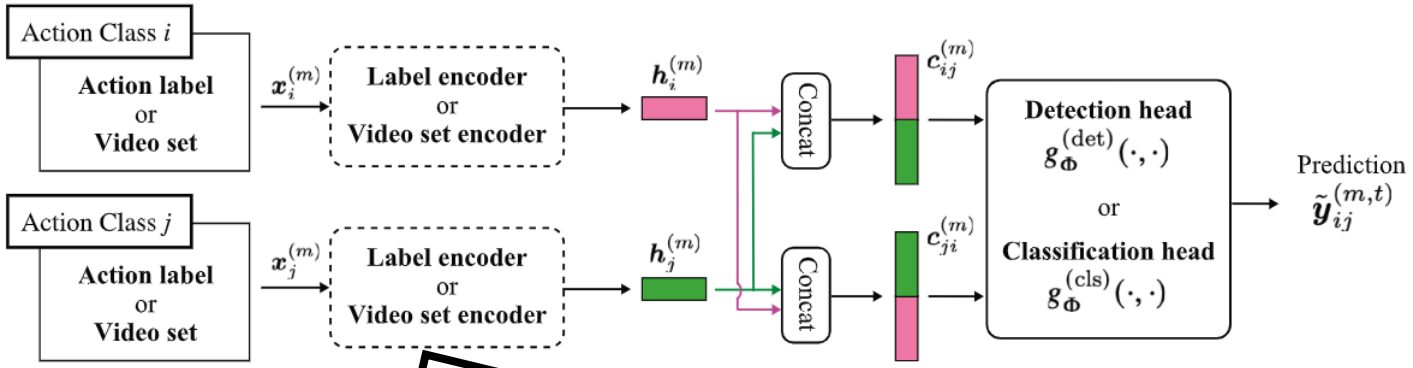
役に立つ場面

- 独自データセットをMetaVDに含めて、独自データセットを拡張
- 新たに既存データセットをMetaVDに追加する際の支援

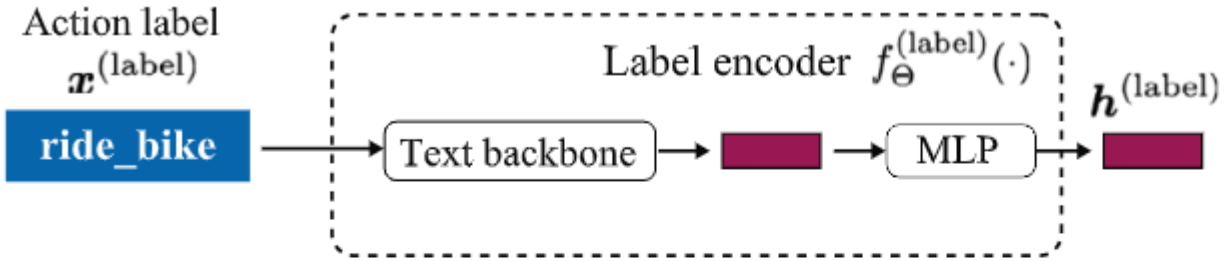
動作ラベルの関係予測 [Yoshikawa+ 2023]

動作クラスi, jのラベル文字列と動画集合から特徴抽出し、
得られた特徴ベクトルからMLPを介して関係の検出・分類を行う

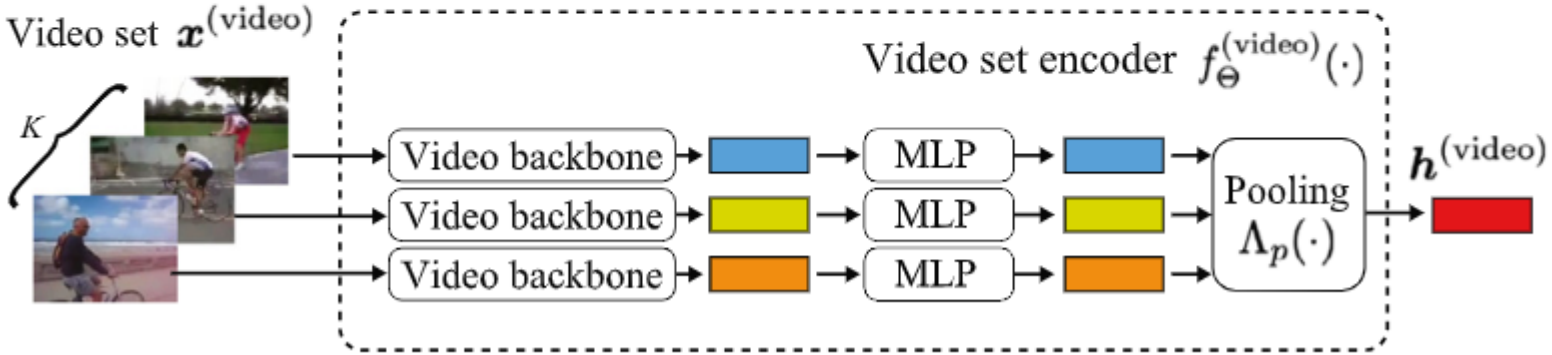




ラベル文字列
エンコーダ
(BERT)



動画集合
エンコーダ
(SlowFast)



動作ラベルの関係予測の性能

UCF101が新たにMetaVDに含めたいデータセットと仮定して、
それ以外の5つのデータセットで関係予測モデルを学習

関係の有無を予測するタスクの予測性能 (Average Precision)

ラベル文字列のみ	動画集合のみ	両方	ランダム
0.711	0.442	0.746	0.006

関係の種類を予測するタスクの予測性能 (Accuracy)

ラベル文字列のみ	動画集合のみ	両方	ランダム
0.785	0.715	0.794	0.522

大規模言語モデル(LLM)で動作ラベル関係予測

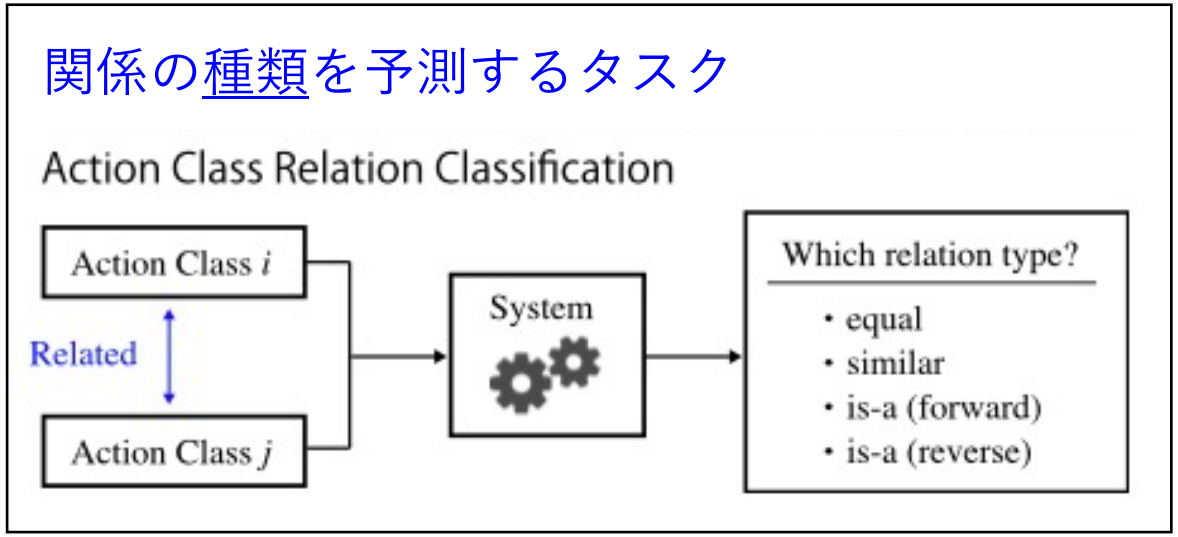
ユーザ

You are an AI that answers the relationship between given two action classes. The relationships between action classes and their definition are as follows:
- "equal": action class 1 and action class 2 are the same meaning.
- "similar": action class 1 and action class 2 are similar meaning.
- "is-a": action class 1 is a superordinate concept of action class 2

Answer the relationship between the following action classes: "eat" and "Having_an_ice_cream"

is-a

関係の種類を予測するタスク



	関係分類精度 (全データセット平均)
GPT-3.5 (finetuned)	**0.914**
GPT-4 (zero-shot)	0.878
GPT-3.5 (zero-shot)	0.565

発展の方向性

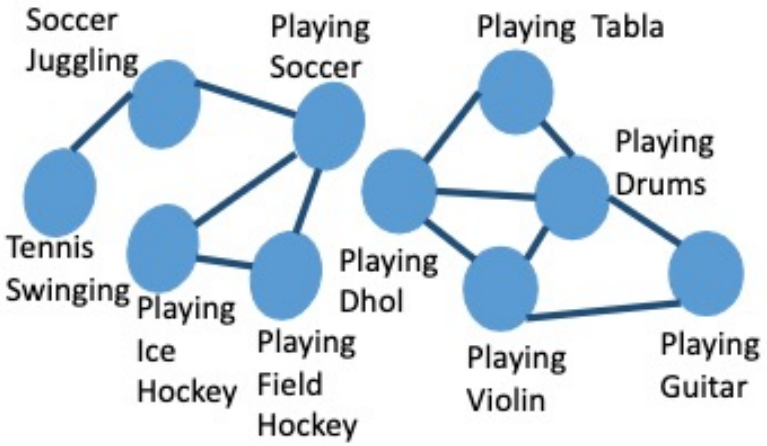
- **深く広い知識の活用**
  - 現状は直接関係する動作ラベルのみを考慮しているが、間接的に関係する動作ラベルの影響も考慮した方がいいのではないか？
  - 動作以外にもその動作と関連する物体等の情報も一緒に扱えるようにした方がいいのではないか？

知識グラフを活用した動作認識

知識グラフを利用したゼロショット動作認識 [Ghosh+ 2020]

UCF101+Kinetics400の動作ラベルの知識グラフを構築

- ノードが動作ラベル（訓練とテストで異なる動作ラベル）
- 各ノードの特徴量は、動作ラベルから得た文埋め込みベクトル
- エッジは特徴量の類似度に基づいて定義



グラフニューラルネット(GNN)で動作ラベルの表現学習

各ノードの最終層が動作認識器の最終層の重みと同じになるように学習

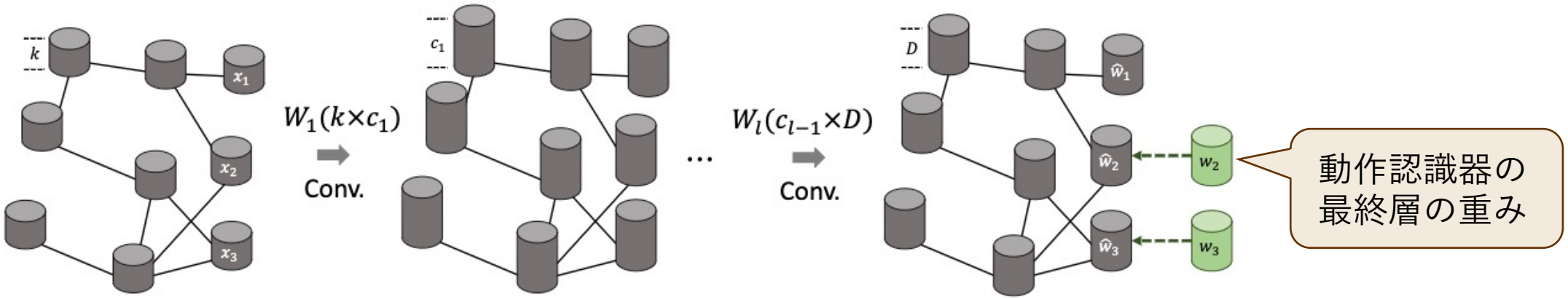
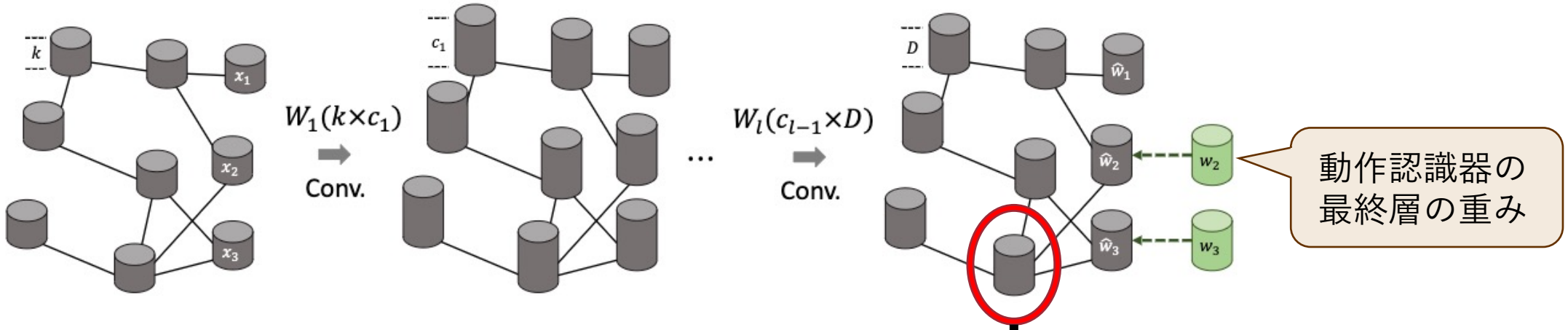


図: [Wang+ 2018]

知識グラフを利用したゼロショット動作認識 [Ghosh+ 2020]

グラフニューラルネット(GNN)で動作ラベルの表現学習

各ノードの最終層が動作認識器の最終層の重みと同じになるように学習

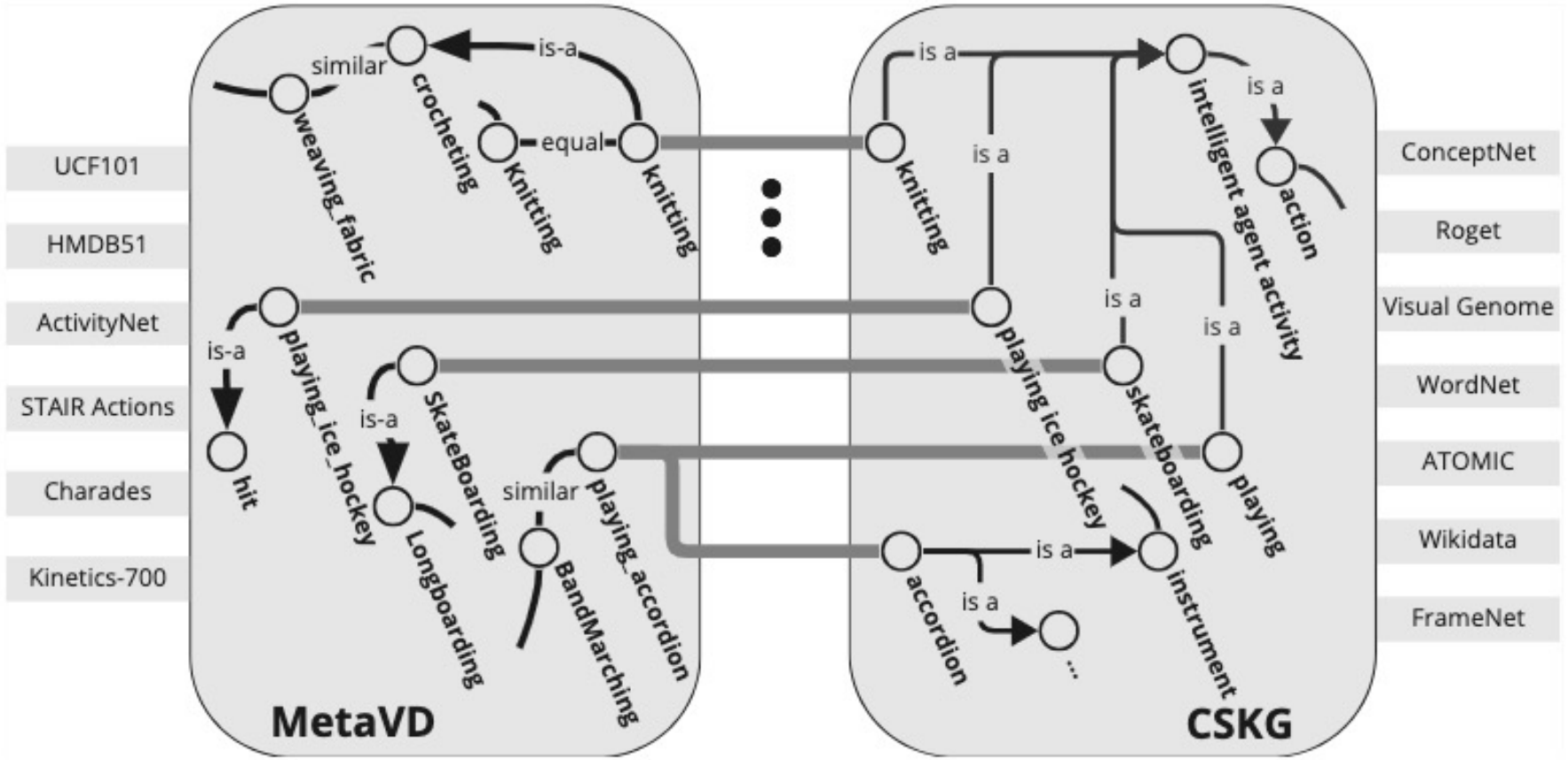


GNNの最終層と動画から得られる特徴量で動作認識

動作認識器にとって未知の動作クラスであっても認識可能

MetaVDを知識グラフに接続 [Yamamoto+ 2023]

知識グラフ (CommonSense Knowledge Graph; CSKG)のノードにMetaVD内の動作を対応付け



CSKGの膨大な知識を活用した動作認識性能の向上が期待できる

発展の方向性

- **深く広い知識の活用**
  - 現状は直接関係する動作ラベルのみを考慮しているが、間接的に関係する動作ラベルの影響も考慮した方がいいのではないか？
  - 動作以外にもその動作と関連する物体等の情報も一緒に扱えるようにした方がいいのではないか？

知識グラフを活用した動作認識

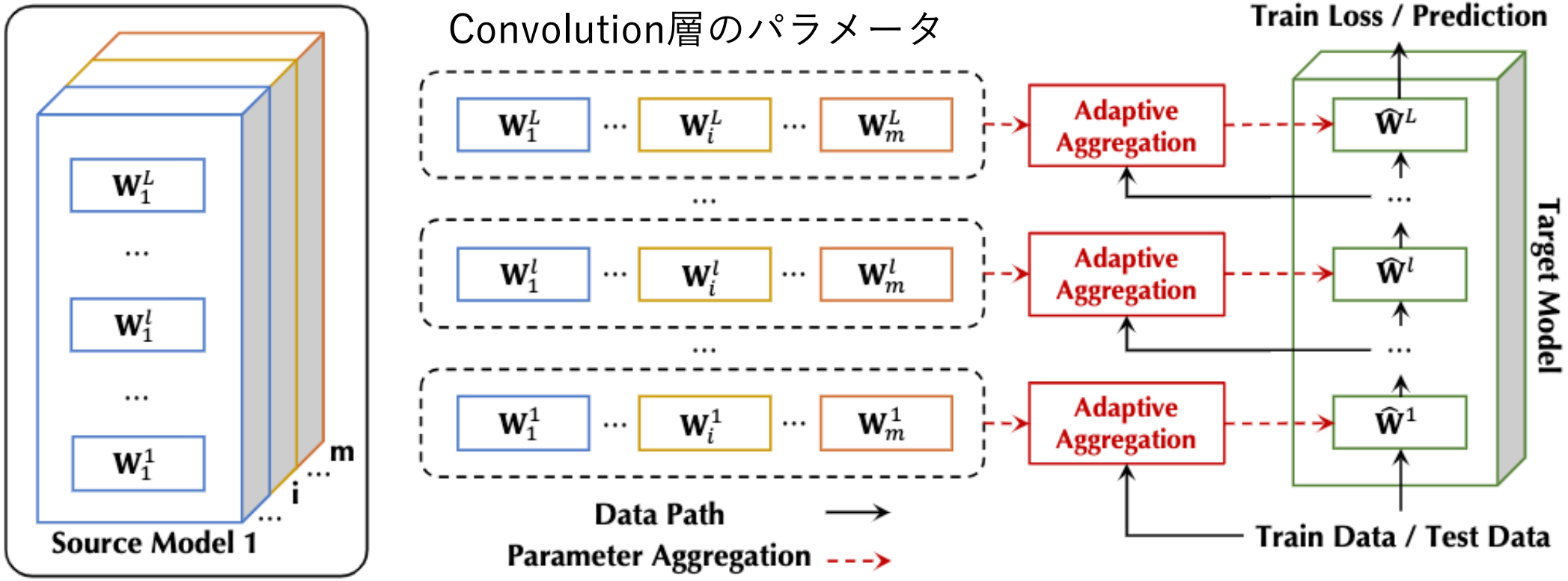
- **訓練コストの削減**
  - 転移元のデータセットではなく、学習済みモデルを利用できれば
    もっと軽量な計算でターゲットデータセットの動作認識器が作れるのではないか？

Model Zooを活用した動作認識

Model Zooからの転移 [Shu+ 2021]

Model Zoo (学習済みモデル置き場)にある異なるデータセットで学習されたモデル (ResNet)をターゲットデータセットのみを用いて転移させる

※ ソース（転移元）モデルもターゲットモデルも同じネットワーク構造を仮定



Model Zooからの転移 [Shu+ 2021]

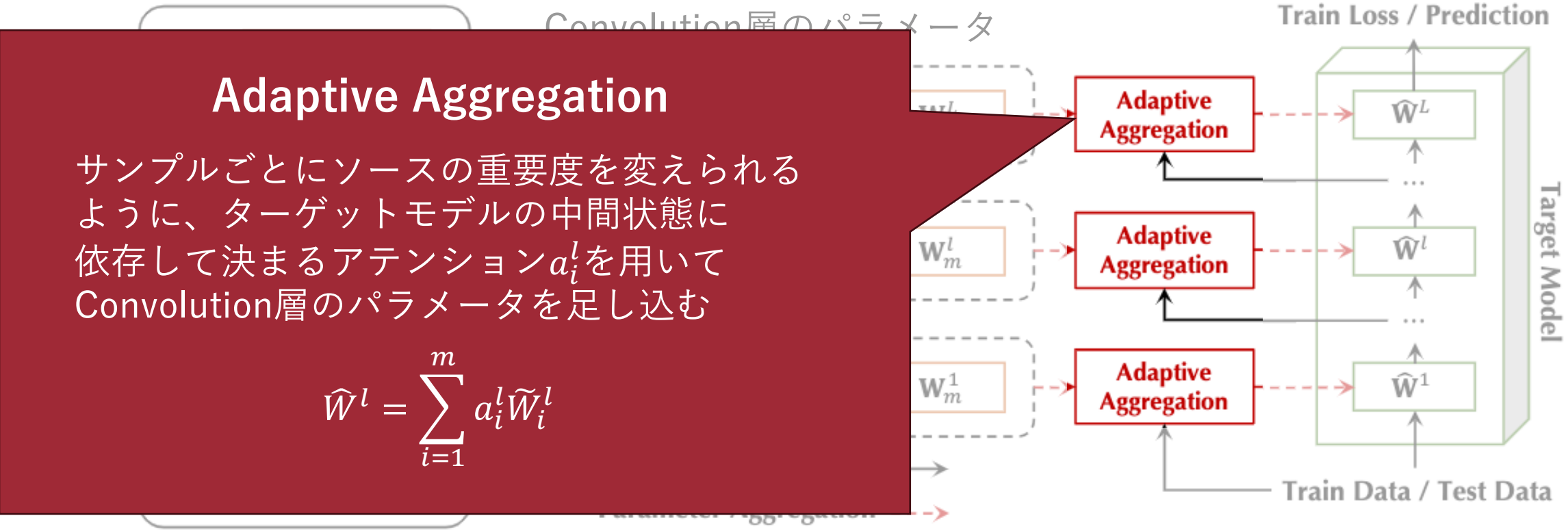
Model Zoo (学習済みモデル置き場)にある異なるデータセットで学習されたモデル (ResNet)をターゲットデータセットのみを用いて転移させる

※ ソース（転移元）モデルもターゲットモデルも同じネットワーク構造を仮定

Adaptive Aggregation

サンプルごとにソースの重要度を変えられるように、ターゲットモデルの中間状態に依存して決まるアテンションa_i^lを用いてConvolution層のパラメータを足し込む

$$\widehat{W}^l = \sum_{i=1}^{m} a_i^l \widetilde{W}_i^l$$



まとめ　「メタ動画データセットによる動作認識の現状と可能性」

- メタ動画データセットによる動作認識の現状
  - 動作認識データセットの動作ラベル間の関係をアノテーションした Meta Video Dataset (MetaVD)を紹介
  - MetaVDを用いたデータセット拡張による動作認識器の学習法
  - 独自データセットをMetaVDに取り込むための関係予測
- MetaVD研究の発展の方向性
  - 知識グラフを活用した動作認識
  - Model Zooを活用した動作認識

MetaVDのダウンロード・可視化ツール・更新情報は以下からアクセスできます

https://metavd.stair.center

参考文献

- [Yoshikawa+ 2021] Yoshikawa, Yuya, et al. “MetaVD: A Meta Video Dataset for Enhancing Human Action Recognition Datasets.” Computer Vision and Image Understanding: CVIU, vol. 212, Nov. 2021, p. 103276.
- [Yoshikawa+ 2023] Yoshikawa, Yuya, et al. “Action Class Relation Detection and Classification across Multiple Video Datasets.” Pattern Recognition Letters, vol. 173, Sept. 2023, pp. 93–100.
- [Yamamoto+ 2023] Yasunori Yamamoto, Shusaku Egami, Yuya Yoshikawa, Ken Fukuda, “Towards Semantic Data Management of Visual Computing Datasets: Increasing Usability of MetaVD,” Proceedings of the ISWC 2023 Posters, Demos and Industry Tracks co-located with 22nd International Semantic Web Conference (ISWC2023), Athens, Greece, Nov. 2023.
- [Ghosh+ 2020] Ghosh, Pallabi, et al. “All About Knowledge Graphs for Actions.” arXiv [cs.CV], 28 Aug. 2020, http://arxiv.org/abs/2008.12432. arXiv.
- [Wang+ 2018] Wang, Xiaolong, et al. “Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, https://doi.org/10.1109/cvpr.2018.00717.
- [Shu+ 2021] Shu, Yang, et al. “Zoo-Tuning: Adaptive Transfer from A Zoo of Models.” Proceedings of the 38th International Conference on Machine Learning, edited by Marina Meila and Tong Zhang, vol. 139, PMLR, 18--24 Jul 2021, pp. 9626–37.