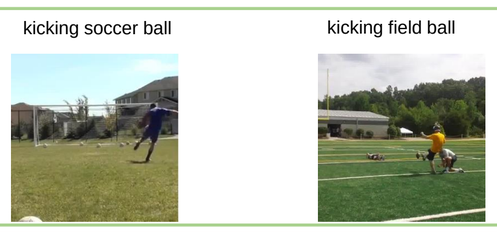
時空間アダプタを用いた動作認識のためのマルチドメイン学習

大見一樹　玉木徹　国立大学法人 名古屋工業大学

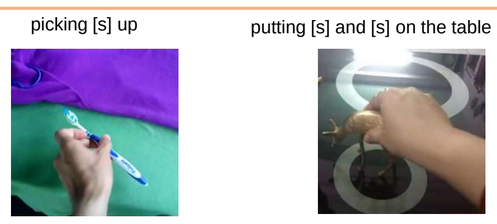
マルチドメイン学習（MDL）

ドメイン
- データが持つ特有の傾向
- 一般的にはドメイン毎にモデルが必要

Kinetics
人物の動作
背景がヒント



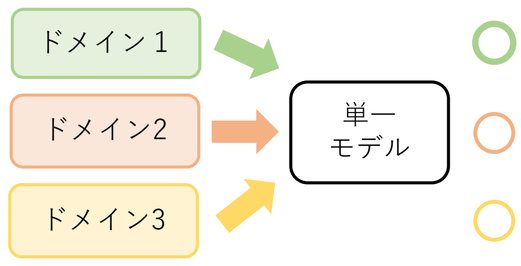
SSv2
人物の動作
時間的な特徴
背景固定



Epic-Kitchens
調理の様子
1人称視点

マルチドメイン学習（MDL）

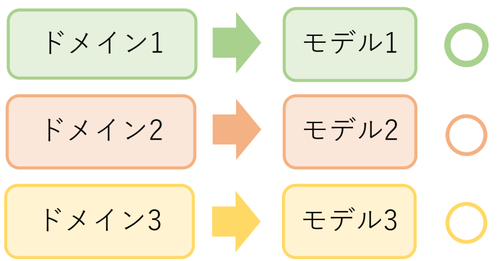
複数のドメインに対して高性能



- 複数のドメインに対応する単一モデルの学習
- ドメイン非依存とドメイン依存の2つを学習
- 動作認識にMDLを適用した研究はない

一般的な学習

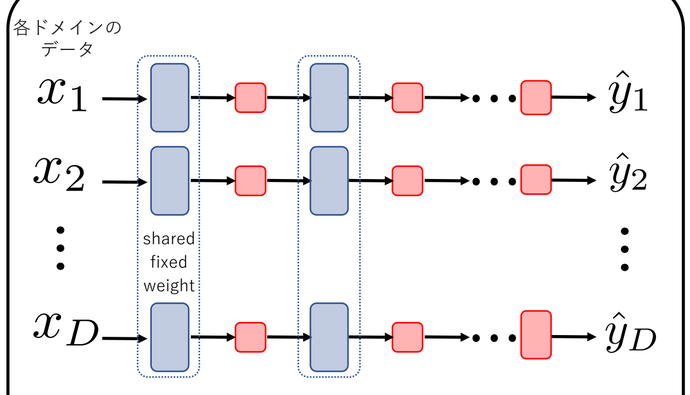
学習したドメインには高性能



学習ドメインと異なると性能劣化

従来手法

アダプタ型

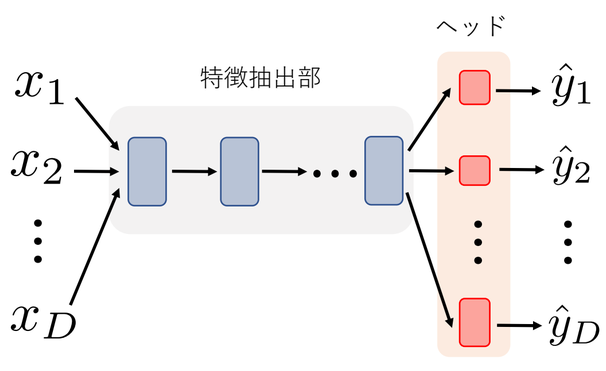


学習方法
- ドメイン非依存は事前学習時の重みで固定
- ドメイン毎にアダプタのみ学習

欠点
- 事前学習時のドメインと異なると性能劣化

従来手法
- Residual adapter [Rebuffi+, NIPS2017, CVPR2018]
- CovNorm [Li & Vasconcelos, CVPR2019]

マルチヘッド型

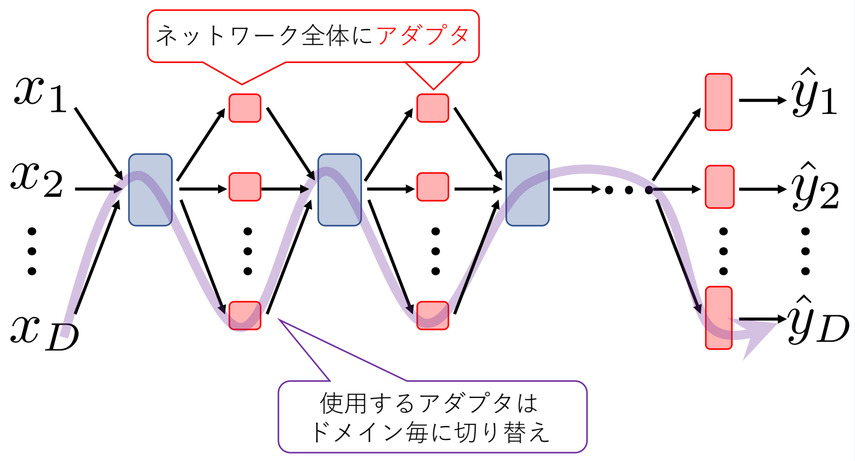


学習方法
- 全てのドメインを同時に学習
- ドメイン毎にヘッドのみを切り替える

欠点
- 特徴抽出部だけで全てのドメインに対応するのは困難

従来手法
- セマンティックセグメンテーション [正木ら, SSII2021]

提案手法



学習方法
- ドメイン毎にヘッドだけでなくアダプタも切り替える
- ドメイン毎に損失を計算し勾配を逆伝播
- 全てのドメインを逆伝播した後にパラメータ更新

利点
- 複数ドメインから普遍的な特徴を学習
- 全てのドメインを同時にend-to-endで学習可能

実験結果

設定

データセット
- UCF101 [Soomro+, arXiv2012]
- HMDB51 [Kuehne+, ICCV2011]
- Kinetics400 [Kay+, arXiv2017]

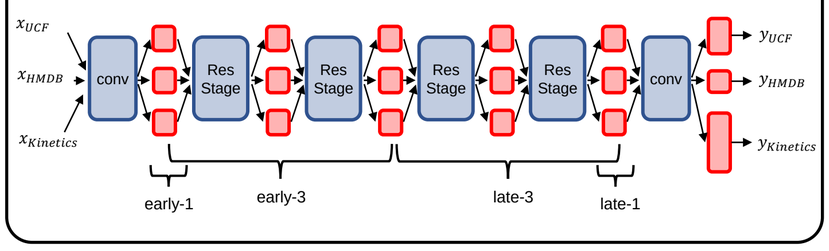
モデル
- バックボーン：X3D-M [Feichtenhofer, CVPR2020]
- Kinetics400で事前学習済み
- アダプタ数：5箇所 × 3ドメイン

提案手法の性能

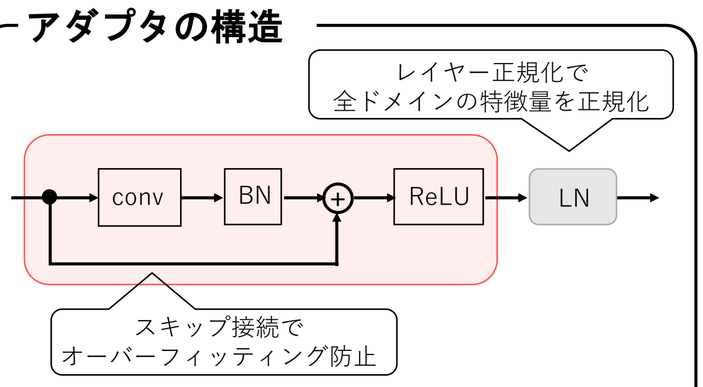
	UCF	HMDB	Kinetics	mean
アダプタ	95.19	73.07	67.54	78.60
マルチヘッド	96.25	73.07	70.62	79.98
提案手法	96.25	74.77	69.89	80.29

- UCFとHMDBで提案手法が最良
- Kineticsはマルチヘッドが良い
  - X3Dが Kineticsで事前学習済みのため
- 平均性能は提案手法が最も高い

アダプタの挿入位置



アダプタの構造



- 畳み込みは(2+1)D convを利用

単一ドメインで学習した場合との比較

	#par.	UCF	HMDB	Kinetics	mean
単一ドメイン	11.9M	96.88	73.27	71.80	80.65
提案手法	5.9M	96.25	74.77	69.89	80.29

- 平均性能を大幅に下げずパラメータ数半減

アダプタの位置と数の影響

	UCF	HMDB	Kinetics	mean
multi head	96.25	73.07	70.62	79.98
early-1	96.38	74.77	71.00	80.72
early-3	96.19	74.64	70.75	80.53
late-1	96.03	73.99	70.86	80.29
late-3	95.90	74.90	70.45	80.42
all	96.25	74.77	69.89	80.29

マルチヘッド型との比較
- UCFは大差なし（-0.35%～+0.13%）
- HMDBは性能向上（+0.92%～+1.83%）
- Kineticsは性能低下（-0.88%～+0.38%）
- 平均性能はどこにアダプタを追加しても向上
  - アダプタは有効