2018年度 卒業論文

ニューラルネットワークを用いた動画像内の物体認識

京都大学総合人間学部認知情報学系

中村優太

2018年 1月 30日 提出

目次

[要旨 4](#_Toc536302862)

[第1章 序論 5](#_Toc536302863)

[第2章 方法 8](#_Toc536302864)

[2.1 畳み込みニューラルネットワーク 8](#_Toc536302865)

[2.1.1 比較に用いた畳み込みニューラルネットワーク 8](#_Toc536302866)

[2.1.2 畳み込みニューラルネットワークの拡張 8](#_Toc536302867)

[2.2 データセット 10](#_Toc536302868)

[2.2.1 Moments In Time データセット 10](#_Toc536302869)

[2.2.2 データセットの抽出 10](#_Toc536302870)

[2.3 物体判別学習 10](#_Toc536302871)

[2.3.1 2次元畳み込みニューラルネットワーク 10](#_Toc536302872)

[2.3.2 3次元畳み込みニューラルネットワーク 11](#_Toc536302873)

[2.3 検証 11](#_Toc536302874)

[2.3.1 評価方法 11](#_Toc536302875)

[2.3.2 評価指標 12](#_Toc536302876)

[3章 結果 13](#_Toc536302877)

[3.1 学習曲線 13](#_Toc536302878)

[3.2判別結果 15](#_Toc536302879)

[第4章 考察 18](#_Toc536302880)

[第5章 結論 20](#_Toc536302881)

[謝辞 21](#_Toc536302882)

[参考文献 22](#_Toc536302883)

# 要旨

ニューラルネットワークのfine-tuningはニューラルネットワークの訓練手法の一つであり, 学習済みニューラルネットワークを元とすることで限られたデータ量でのタスクの学習を可能にするため, 近年注目を集めている技術である. 特に, ニューラルネットワークが注目されるきっかけともなった静止画を扱うタスクにおいては, fine-tuning が一般的に用いられており, その方法論も確立されている. 一方で動画を扱うタスクにおけるfine-tuningの方法論は確立されておらず, タスクによって適した手法でfine-tuningを行うことによりニューラルネットワークの性能が向上する可能性がある. そこで本研究では動画中の物体判別タスクを対象として, ネットワークのアーキテクチャと学習済みタスクが異なるニューラルネットワークを用いてfine-tuningを行い, 学習後のニューラルネットワークによる物体判別タスクの成績の比較を行うことでfine-tuningに用いるべき学習済みニューラルネットワークについて検証した. 検証の結果, 動画用の構造を持つネットワークに静止画中の物体判別タスクを学習させたネットワークを用いてfine-tuningを行うことにより動画中の物体判別タスクを学習できることが示された. これは, 限られたデータ量であってもfine-tuningの元とするネットワークを精査することで動画を扱うタスクの学習を行えることを示唆している.

# 第1章 序論

ニューラルネットワークは大規模なデータベースを用いることによって, 多様な課題において革新的な性能の向上をもたらしてきた. 大量のデータの学習により, 画像認識・音声認識・自然言語処理など様々な課題においてこれまでにない成果を出してきた. ニューラルネットワークが注目されると共に, 機械学習に用いられるデータセットの大型化が進み百万件以上のデータを含むデータベースの使用も一般的なものとなった.

一方で, データの量が限られている状況でニューラルネットワークを訓練する技術も研究されてきた. その一例としてfine-tuningが挙げられる. Fine-tuningは, あるタスクのために学習されたニューラルネットワークの重みを初期値として用いることにより, 別のタスクの学習を行う手法である. Fine-tuningを用いてニューラルネットワークを訓練する際には, 学習の第一段階としてターゲットとするタスクとは異なる大量のデータを用意できるタスクを学習し, その後にデータ量が限られているターゲットとなるタスクを行うように訓練を行う. その際に始めに行うタスクの学習を通して, ニューラルネットワークがfine-tuningのターゲットとなるタスクにおいて有用となる特徴を抽出している必要がある.

一般に, 静止画像識別の領域においては静止画中の物体判別タスクを学習したニューラルネットワークがこのような特徴量を抽出しているとされ,静止画中の物体判別タスクを学習したネットワークが一般的にfine-tuningの際の元のモデルとして使用されている. この方法はセグメンテーションやキャプションの生成など多くの画像認識タスクにおいて成功を収めている (Shelhamer, Long, & Darrell, 2017; Vinyals, Toshev, Bengio, & Erhan, 2015).

また, fine-tuningは静止画を扱うタスクだけではなく, 動画を扱うタスクにおいても効果的であることが示され始めている. 一例として動画中の動詞判別タスクにおいて, 大量のデータを有するkineticsデータセット (Kay et al., 2017) で学習したモデルを元として, より小規模なデータセットにおける動詞判別タスクを学習することが可能であることが示されている (Carreira & Zisserman, 2017) . しかし, 動画を扱うタスクのfine-tuningの手法は現状ではまだ確立されていない. 静止画認識の分野においてはpre-trainingタスクとして静止画中の物体判別タスクが一般的に用いられているが, 動画認識の分野においては, 動画中の動詞判別タスクや静止画中の物体判別タスクなど複数の pre-trainingタスクの候補が存在し, 動画認識のタスクを学習する際に, どういった手法をfine-tuningの手法として用いるのが良いかは明らかになっていない.

そこで, 本研究では動画中の物体判別タスクを学習するために, どうようなfine-tuningの手法を用いるのが効果的かを検証した. ニューラルネットワークのアーキテクチャとpre-trainingとして学習するタスクを操作することにより複数の学習済みニューラルネットワークを用意し, それらを同様の方法でfine-tuningを行い動画中の物体判別タスクの成績を比較することで, 動画中の物体判別タスクを学習するために最適なfine-tuningの手法を検証した.

第2章の第1節では, 今回の検証に使用したニューラルネットワークについての詳細を解説する. 第2 章の2, 3 節では, 今回の検証に用いたデータセットと学習の際の手続きについて述べる. 第2 章の4 節では, 比較検討の方法論について述べた後, 第3 , 4 章では, 比較の結果および検証を行い, ニューラルネットワークのfine-tuningについて考察を行う.

|  |  |
| --- | --- |
| ネットワークアーキテクチャ | 学習済みタスク |
| 2次元畳み込み (静止画用) | 静止画中の物体判別タスク |
| 3次元畳み込み (動画用) | 静止画中の物体判別タスク |
| 3次元畳み込み (動画用) | 動画中の動詞判別タスク |

**図 1 ．検証に用いた学習済みニューラルネットワーク.**  Fine-tuning手法の検証のために用いた学習済みのニューラルネットワーク. ニューラルネットワークのアーキテクチャと, 学習に用いたタスクの組み合わせを変動させることにより学習済みモデルを作成した.

# 第2章 方法

## 2.1 畳み込みニューラルネットワーク

本検証では, ニューラルネットワークとして, 画像認識・動画認識の分野において一般的に用いられている畳み込みニューラルネットワークを使用した. 畳み込みニューラルネットワークは畳み込み層やプーリング層を重ね合わせることで構成されるニューラルネットワークであり, 入力とするデータの次元によって異なる構造のものを用いる. 本研究では, 静止画像の入力を前提とした2次元の畳み込みを行うネットワークである2次元畳み込みニューラルネットワークと動画の入力を前提とした3次元の畳み込みを行う3次元畳み込みニューラルネットワークを使用して検証を行った.

### 2.1.1 比較に用いた畳み込みニューラルネットワーク

本検証では, 訓練済み畳み込みニューラルネットークをfine-tuningし, 物体判別タスクの検証を行った. 2 次元の畳み込みニューラルネットワークで画像中の物体判別タスクを行うネットワークとして, ImageNet (Jia Deng et al., 2009) を用いた1000クラス物体判別タスクでpre-trainingされた ResNets (He, Zhang, Ren, & Sun, 2016) を用いた. 本検証においては, 50層のResNetsを用いた.

また, 3 次元の畳み込みニューラルネットワークとして, 画像中の物体判別タスクでpre-trainingされた2 次元の畳み込みニューラルネットワークを拡張した3次元畳み込みニューラルネットワークと, 同様のネットワークを動画中の動詞判別でpre-trainingしたネットワークを用いた.

### 2.1.2 畳み込みニューラルネットワークの拡張

時空間方向の畳み込みを行う3 次元の畳み込みニューラルネットワークで, 静止画中の物体判別タスクを行うものとして, I3D ネットワーク (Carreira & Zisserman, 2017) を用いた．I3Dネットワークは, 訓練済みの2 次元畳み込みニューラルネットワークを3次元に拡張することにより作られる3次元畳み込みニューラルネットワークである. 拡張は3次元畳み込みニューラルネットワークの作成と, 2次元畳み込みニューラルネットワークからの学習済みの重みの転移によって行われる. 3次元畳み込みニューラルネットワークは, 畳み込みニューラルネットワークの畳み込み層とプーリング層に時間方向の次元を加えることにより作成される. ネットワークを作成した後の重みの転移は, 3次元畳み込みニューラルネットワークに2 次元の同じ画像を繰り返すことで作成された動きがない動画 (boring-video) を入力した時の出力が, もとの2 次元畳み込みニューラルネットワークに同じ画像を入力した時の出力と等しくなるような制約を満たすように行う.

本検証では二つの方法で上記の制約を満たす拡張を行った. それぞれの方法において3次元畳み込みニューラルネットワークの畳み込み層の重みは変換前の2次元畳み込みニューラルネットワークにおいて対応する畳み込み層の重みから転移を行った. 一つ目の手法では, 変換する層の時間軸方向の大きさが*N*のとき, 対応する畳み込み層を重みを保ったまま時間方向に*N*回重ねた後に, 1/*N*倍することにより時空間ネットワークの重みの初期化を行った. 二つ目の方法では, 3次元畳み込みニューラルネットワークの重みをすべて0で初期化を行った後に, 時間軸において中央に位置するフィルターにのみ対応する2 次元畳み込みニューラルネットワークの重みを転移することによって初期化を行った. 本研究においては, 前者を平均化拡張 (Carreira & Zisserman, 2017) , 後者を中心化拡張 (Girdhar et al,. 2018) と呼ぶ.

本検証においては, 画像中の物体判別にpre-trainingされた3次元畳み込みニューラルネットワークとして, ImageNetでpre-trainingされたResNets50 をそれぞれ, 平均化拡張, 中心化拡張によって3次元畳み込みニューラルネットワークに拡張したものを用いた.

#### 2.1.2.2 動詞判別3次元畳み込みニューラルネットワーク

3次元畳み込みニューラルネットワークで, 動画中の動詞判別を行うニューラルネットワークとして kineticsデータセットの動詞判別でpre-trainingされたニューラルネットワークを用いた. このニューラルネットワークは前述のImageNetでpre-trainingしたI3Dネットワークを元にkineticsデータセットでの動詞判別のタスク用にfine-tuneされたものであり, ネットワークの構造としては前述のものと同様のResNets50を使用しているものを用いた.

## 2.2 データセット

### 2.2.1 Moments In Time データセット

I3Dの訓練, および検証にはMomets In Timeデータセット (Monfort et al., 2018) から抽出した1200 件の動画データ及び, 動画に対応する物体ラベルラベルを使用した. Moments In Timeデータセットは100万枚以上の3秒間の動画に339種類のアクションのラベルが動詞名で一つずつ付けられたデータセットであり, 同様のものとしては最大規模のデータセットである.

### 2.2.2 データセットの抽出

本研究では, 動詞ラベルではなく動画中の物体ラベルラベルを利用するため, Moments In Timeデータセットから訓練, テスト用のデータとして150種類のアクションラベルがついた動画をそれぞれ8件とした1200件のデータを元としてデータセットを作成した.これらの動画に対するラベリングを, それぞれの動画中に確認できる物体のラベルを複数つける形で行った. ラベリングを行った結果, 193の物体ラベルがラベルとして与えられ, 1動画あたりの平均ラベル数は1.41であった. 本研究では, このうち出現頻度上位20ラベルのみを抽出して用いた. 抽出され,検証に用いられた動画は937件, 1動画あたりの平均のラベル数は1.25であった. ラベルが付けられた動画データは, 全て時間が3秒間, フレーム数90枚, 解像度は縦256画素, 横256画素であった. 本検証においては, 90フレームの動画から1フレームごとにフレームを抽出し45フレームの動画とし, その中央32フレームを抽出して作成したデータを学習・検証に用いた.

## 2.3 物体判別学習

### 2.3.1 2次元畳み込みニューラルネットワーク

2 次元の畳み込みニューラルネットワークは以下の方法で訓練を行った. ニューラルネットワークへの入力は, 作成したデータセット中の動画データの32フレームをそれぞれ一枚の画像とし, 全動画の全フレームをランダムにシャッフルした後, 16枚を1バッチとして行った. また, それぞれの入力画像に対して, 左右, 上下の反転をおこなった後, 256 × 256の解像度の画像から224 × 224の解像度の画像をランダムな位置で切り抜く前処理を行った.

また, 学習時の条件は以下のものを用いた. 損失関数には最終層の出力にシグモイド関数を適用した各ラベルの予測値と, 真のラベルとのクロスエントロピーの全ラベル間での平均を用いた. 最適化手法としては, Momentum stochastic gradient descent 法 (Momentum SGD) を, Momentumの値を0.9として使用した. 4バッチ毎に勾配を蓄積し, その勾配を用いて重みを更新した. 本研究においては, 一回の重みの更新を1ステップと呼ぶ. 学習率は初期値として0.01を用い, それぞれ300ステップ, 1000ステップの学習後に0.1倍した. また, 学習の際はWeight Decay を用いた重みの正則化を行った.

### 2.3.2 3次元畳み込みニューラルネットワーク

3次元の畳み込みニューラルネットワークは以下の方法で訓練を行った. ニューラルネットワークへの入力は, 抽出した32フレームの動画1つを1バッチとして入力を行った. また, それぞれの入力動画に対して, 左右, 上下の反転をおこなった後, 256 × 256の解像度の動画から224 × 224の解像度の動画をランダムな位置で切り抜く前処理を行った. また, 学習時の条件は上述の2 次元畳み込みニューラルネットワークと同様のものを用いた.

## 2.3 検証

### 2.3.1 評価方法

畳み込みニューラルネットワークの比較は, fine-tuningを行ったモデルによる動画中の物体判別タスクによって行った. 物体判別タスクを行う際には作成したデータセットのテストデータを用いて行った. 畳み込みニューラルネットワークへの入力は256 × 256 の解像度の画像及び動画から中央の224 × 224を切り抜いたものを使用した. 畳み込みニューラルネットワークの最終層の値を, それぞれの対応するラベルの予測値として評価を行った.

### 2.3.2 評価指標

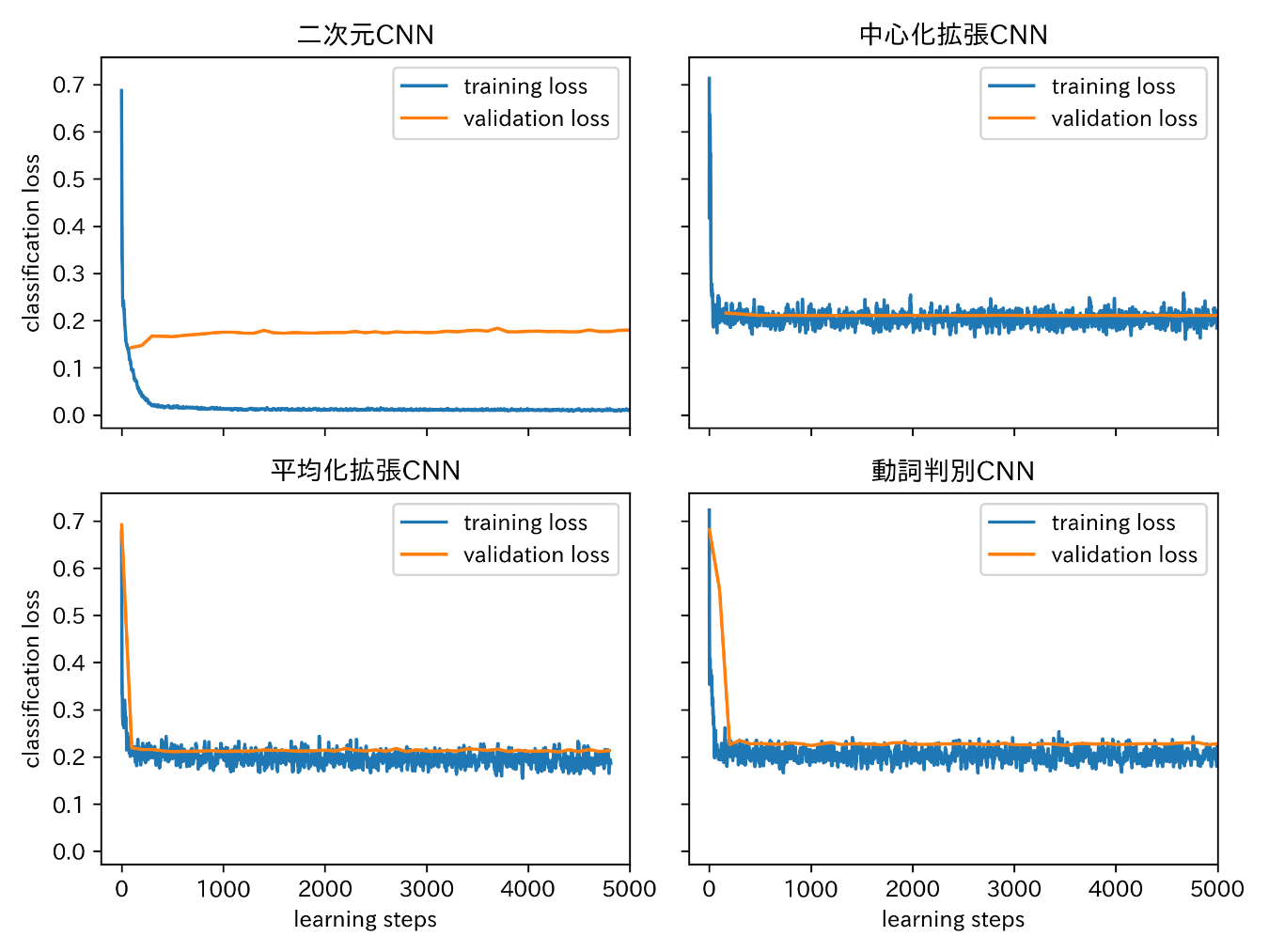
ニューラルネットワークによる予測の評価は, それぞれのラベルの物体が予測画像に含まれているかの二値判別としてArea Under the Curve (AUC) を用いて行った. 20ラベルそれぞれについて, テストセットに対するモデルの予測値と真のラベルを用いてAUCを算出した.

# 3章 結果

Moments In Timeデータセットを用いて, 動画中の物体判別問題における2 次元畳み込みニューラルネットワークと3次元畳み込みニューラルネットワークの動画中の物体判別タスクにおける成績の評価を行った.

## 3.1 学習曲線

Fine-tuningをする際の学習曲線の比較を行った. 損失は最終層の各ラベル毎の予測における交差エントロピー誤差を各ラベルにおいて平均することで求めた. 全ての畳み込みニューラルネットワークにおいて学習初期に急激に損失が減少した後に学習が収束した. 2 次元畳み込みニューラルネットワークにおいてのみ, 訓練データでの損失とテストデータでの損失に大きな差が見られ, それ以外の畳み込みニューラルネットワークにおいては, 訓練データとテストデータにおいて損失の値は大きな差は見られなかった.

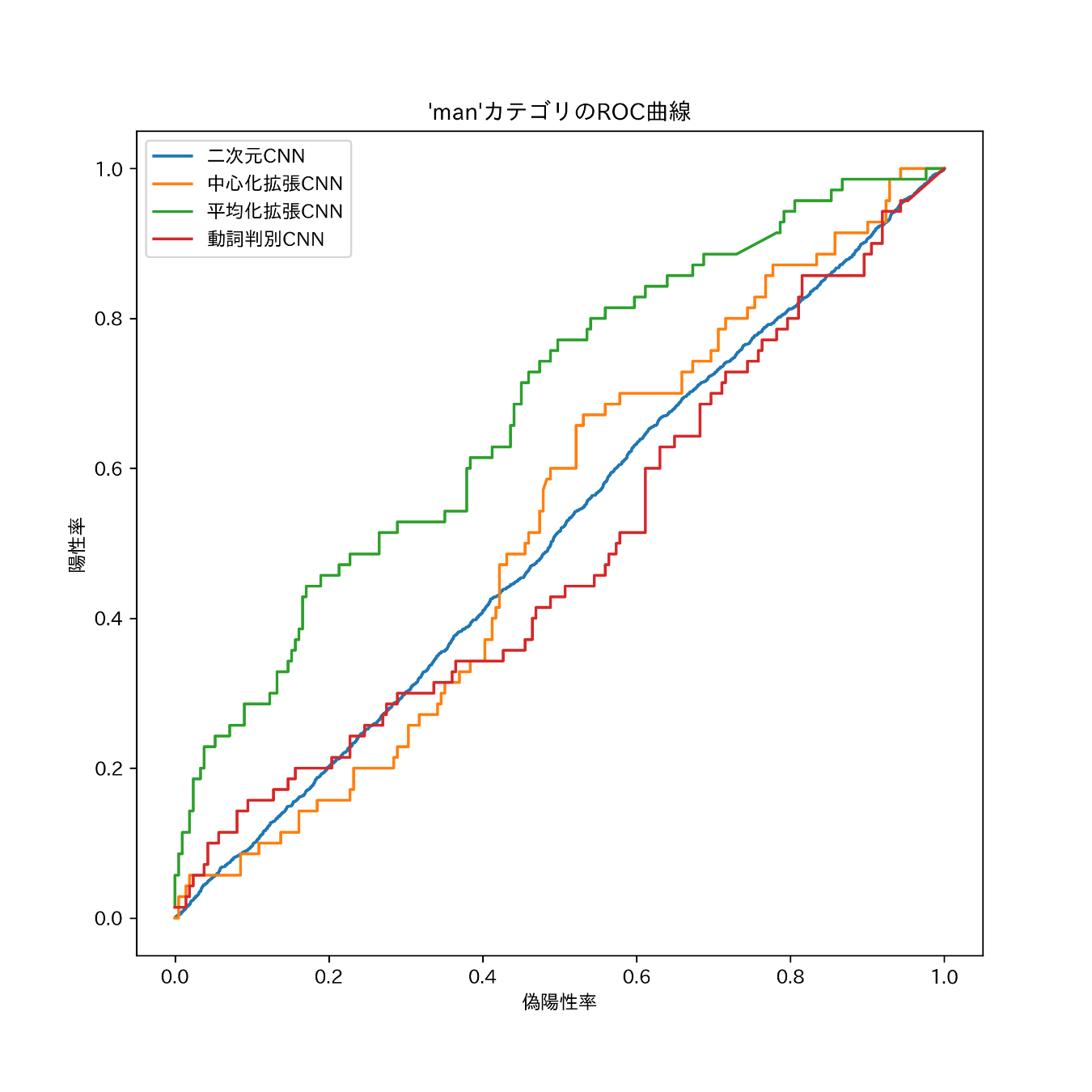


**図 2 畳み込みニューラルネットワークの学習曲線** それぞれの畳込みニューラルネットワークの訓練時の学習時ロスの比較. 左上が2次元畳み込みニューラルネットワーク, 右上が中心化拡張ニューラルネットワーク, 左下が平均化拡張ニューラルネットワーク, 右下が動詞判別畳み込みニューラルネットワークをそれぞれ表している. それぞれに対して, 訓練データに対する損失とテストデータに対する損失を図示している.

.

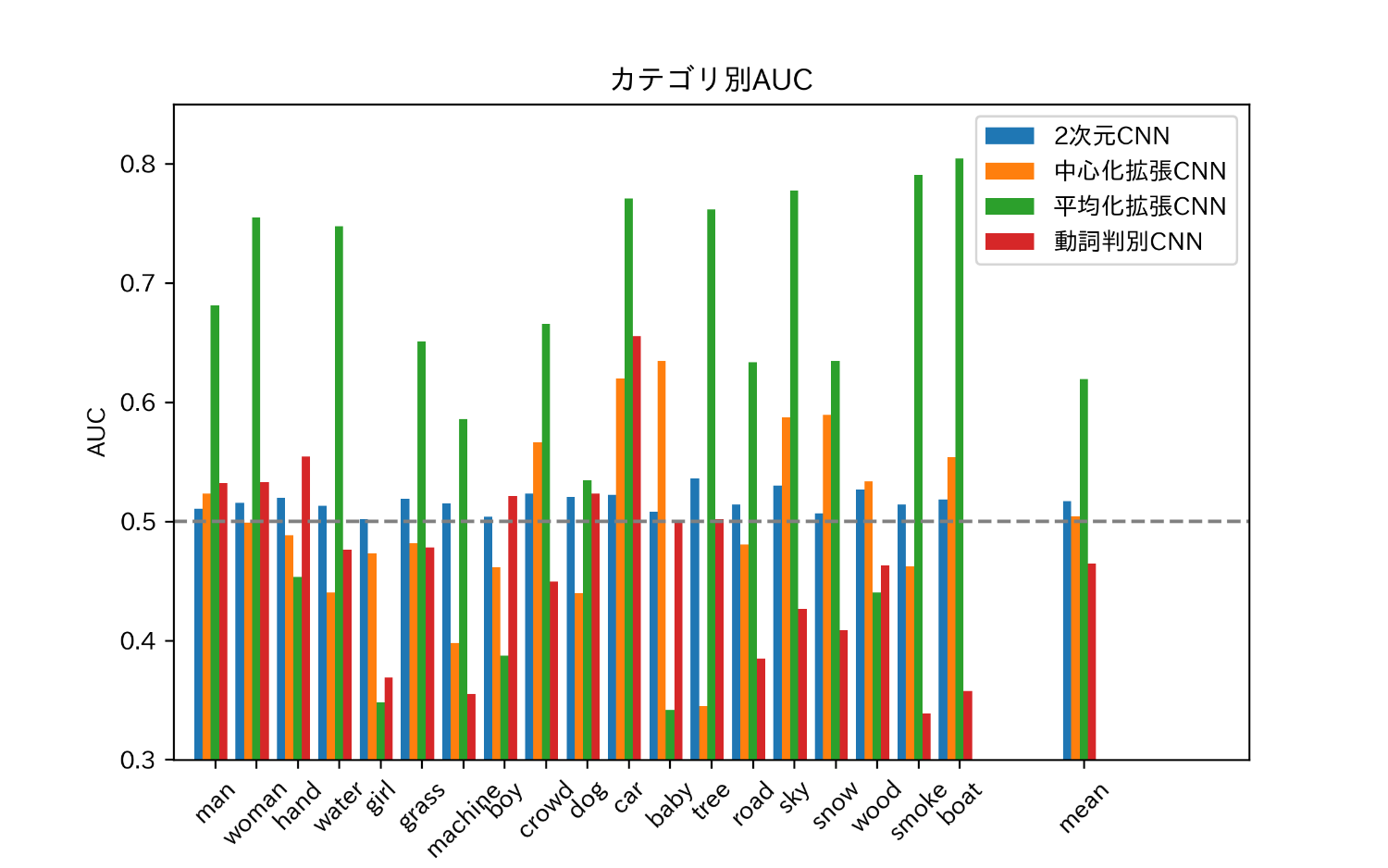
### 3.2判別結果

動画中の物体判別問題の結果の比較を行った. 図3は, いくつかのラベルに対する予測のROC曲線を比較したものである. 中心化拡張によって拡張された3次元畳み込みニューラルネットワーク, 動詞判別3次元畳み込みニューラルネットワーク, 2 次元畳み込みニューラルネットワークにおいてはROC曲線はチャンスレベルのものと同等の結果を示した. 一方で, 平均化拡張によって拡張された3次元畳み込みニューラルネットワークは, 他の畳み込みニューラルネットワークよりも判別成績がよいことが分かる.



**図 3 ラベル判別のROC曲線.** 各畳み込みニューラルネットワークを元にfine-tuningしたネットワークによる “man” ラベルの二値予測に対するROC曲線.

また, 図4は畳み込みニューラルネットワークの予測値から算出した, ラベルごとのAUCである. 画像識別ニューラルネットワークによる判別結果では, 各ラベルのAUCはおおよそ0.5とチャンスレベルと同等となっており, 学習に失敗していることが分かる. また, 中心化拡張によって拡張された3次元畳み込みニューラルネットワークにおいては, 画像識別ニューラルネットワークよりも値の変動が大きいものの, おおよそAUCは0.5付近の値を取っており, 学習は成功していないことが分かる. 平均化拡張によって拡張された3次元畳み込みニューラルネットワークにおいては, 前述の2つのネットワークよりも高いAUCを示しており, 学習が一定成功していることがわかる. 動詞判別3次元畳み込みニューラルネットワークに関しては他のものよりも総合的にAUCが低い結果となった.



**図 4 ラベル毎のAUC比較.** ラベル毎のAUCをそれぞれの畳み込みニューラルネットワーク毎に比較している. 縦軸はAUCの値. ラベルのラベルは左からデータセット中での出現頻度が高いものから順に並んでいる. 一番右側にある mean ラベルは, 各畳み込みニューラルネットワークのAUCのラベル平均を示している.

# 第4章 考察

本研究では, 複数の方法で重みを設定した畳み込みニューラルネットワークのfine-tuningを行い, 動画中の物体判別タスクの成績の比較を行った結果, 今回比較を行った4つの畳み込みニューラルネットワークの中で, 平均化拡張によって2 次元画像識別タスクで訓練を行ったもののみが他のものと比べ高い成績を示すことが分かった.

また, その他の3つのネットワークに関しては, ROC曲線やAUCの値は総合的に見てチャンスレベルに近く, 物体判別タスクの学習は成功しなかった. しかし, 図2 の学習曲線から, 学習が成功しなかった原因が2 次元の畳み込みニューラルネットワークと3次元畳み込みニューラルネットワークにおいて異なることが明らかになった.

2 次元の畳み込みニューラルネットワークからfine-tuningを行った場合は, 訓練データに対する判別誤差が, テストデータに対する判別成績を大きく下回る過学習が起きることが分かった. これは, 動画中のフレームを画像として切り出して訓練を行う際に, 画像としての類似度が非常に高い画像が複数入力されるという特徴によって引き起こされていると考えられる. 一方, 3次元畳み込みニューラルネットワークからfine-tuningを行った場合は訓練データに対する判別誤差とテストデータに対する判別誤差の間の乖離は起きず, 双方とも判別誤差が初期段階で一定となることが明らかになった.

3次元畳み込みニューラルネットワークの中で平均化拡張で拡張された画像識別畳み込みニューラルネットワークのみが物体判別タスクにおいて成績が高かった原因としては, 平均化拡張においては初期の畳み込みニューラルネットワークの重みからの変化量が小さくても, 新しいタスクの学習が行えるという可能性が考えられる. すでに物体判別のタスクで学習されているネットワークを拡張する場合, 2章で述べた中心化拡張と平均化拡張という2つの手法を用いることができるが, 前者の中心化拡張の場合は畳み込み層の重みの大部分の値が 0 という状態から訓練を行う必要があるため, 本検証のようにデータ量が限られている条件においては十分に重みを更新できなかった可能性がある. また, 動詞判別タスクで訓練された3次元畳み込みニューラルネットワークを用いた場合に関しても, 画像判別とは別のタスクで訓練されていたため, 今回のデータ量では十分に重みが変化しなかった可能性も考えられる. 本検証においては, 学習済みのニューラルネットワークの重みについては定量的な評価が行えていないため, 今後の課題として, 学習済み畳み込みニューラルネットワークの重みの分析を行う必要がある.

以上のような比較から, 3次元畳み込みニューラルネットワークにおけるfine-tuningにおいては以下のような特性があると考えられる. まず, 動画を扱う畳み込みニューラルネットワークとしては2 次元の畳み込みニューラルネットワークと3次元畳み込みニューラルネットワークが挙げられるが, 本検証に用いた比較的小規模のデータ量を用いた場合には2 次元の畳み込みニューラルネットワークは過学習に陥る傾向がある. 一方, 3次元畳み込みニューラルネットワークにおいては, 訓練データを含め判別成績が向上しにくいという問題がある.

本研究では, 動画を用いたタスクにおける前述した問題を緩和する方法として, 同様のタスクで訓練された2 次元畳み込みニューラルネットワークを平均化拡張によって3次元畳み込みニューラルネットワークに拡張したネットワークを初期値として用いたfine-tuningが有用であることが示唆された. 今後の課題としては, fine-tuningを行った後の畳み込みニューラルネットワークの学習済みの重みの定量的な分析を行い, 平均化拡張のみが好成績を残したメカニズムを検証することが挙げられる.

# 第5章 結論

本研究では, 動画中の物体判別タスクの学習のための動画のfine-tuningの特性を調査するために, 異なる学習済み畳み込みニューラルネットワークを用いてfine-tuningを行い, 判別結果を比較した. その結果, 動画中の物体判別タスクのfine-tuningにおいては, 同様の画像判別タスクで学習済みの2 次元畳み込みニューラルネットワークを平均化拡張を用いて3次元畳み込みニューラルネットワークに拡張したネットワークを元としてfine-tuningを行うことでタスクの学習に成功するという結果が得られた. これは少量のデータを用いた動画認識タスクのために畳み込みニューラルネットワークをfine-tuningする場合においては, データ量が多い静止画においてターゲットとするタスクに類似するタスクを学習し, それを平均化拡張によって拡張した後にfine-tuningを行う方法が優れていることが示唆している.

# 謝辞

本研究を⾏うにあたり，脳情報学研究室の神⾕之康教授，間島慶助教には数々のご指導, ご協力を頂きました. 研究のみならず, 多岐に渡ってご支援頂いたことに心より感謝しております. ATR脳情報研究所の堀川友慈主任研究員には，論文の推敲にお世話になり感謝いたします. ATR 脳情報研究所の塚本光昭研究技術員には，研究室の計算機環境の構築および研究を円滑に進める上での数々のサポートをしていただき感謝いたします．京都⼤学情報学研究科修⼠課程1 回の白川健さんには, 対象データの準備や解析方法のサポートをしていただきました. 最後に研究に対して⽀援してくださった脳情報学研究室，ATR 脳情報研究所の皆様に感謝いたします．

# 参考文献

Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. In Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672–2680).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., … others. (2017). The kinetics human action video dataset. ArXiv Preprint ArXiv:1705.06950.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., … others. (2018). Moments in time dataset: one million videos for event understanding. ArXiv Preprint ArXiv:1801.03150.

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), 640–651.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489–4497).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156–3164).

Wu, S., Zhong, S., & Liu, Y. (2017). Deep residual learning for image steganalysis. Multimedia Tools and Applications, 1–17.

Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., & Tran, D. (2018). Detect-and-Track: Efficient Pose Estimation in Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 350-359).