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 12 月 28 日 第一稿

# 目次

# 第 1 章 序論

畳み込みニューラルネットワークはニューラルネットワークの一種であり, 主に画像の判別タスクなどにおいて飛躍的な性能の向上をもたらした技術である. 近年の研究においては画像の判別だけではなく, 画像のキャプション生成や [Vinyals, Toshev, …, 2015], 画像生成のタスク [Goodfellow, Pouget-Abadie, …, 2014]においても, これまでにない成果を挙げている. また, 画像だけではなく動画の判別タスクにおいても好成績を収めている [Tran, Bourdev, Fergus, Torresani, Paluri, 2015] . 一方で, 畳み込みニューラルネットワークの学習には大量のデータが必要であることから, 限られたデータでネットワークを学習するための手法が模索されてきた.

fine-tuning は畳み込みニューラルネットワークの学習に用いられる手法であり, あるタスクのために学習された畳み込み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. 通常の畳み込みニューラルネットワークの学習には大量のデータが必要となるが, fine-tuning を用いる場合, 比較的小規模のデータでも学習を行えるため, 特に計算機による画像識別タスクにおいて一般的に用いられている技術である.

fine-tuning を行う際には, 元となるタスクの学習を通して, 畳み込みニューラルネットワークが fine-tuning の対象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, 画像識別の領域においては画像識別タスクを学習した学習モデルがこのような特徴量を抽出しているとされ, 最も一般的に fine-tuning の際の元のモデルとして使用されており, 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果を上げている.

また, fine-tuning の有用性は 2 次元画像の識別だけではなく, 動画の認識においても効果的であることが示されている. 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, Carreira, Simonyan, 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] . 一方で, 3 次元の畳み込みニューラルネットワークにおいては, 動詞判別タスクを学習した畳み込みニューラルネットワークの fine-tuning によって, 別のタスクを学習できるかは十分に知られていない.

また, 動画よりも多くのデータセットや先行研究が存在する画像認識において学習された畳み込みニューラルネットワークとの比較が行われておらず, 動画を対象とした動詞判別タスク以外のタスクを fine-tuning を用いて行う際に元とする学習済み畳み込みニューラルネットワークの選択方法は十分に研究されていないのが現状である.

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, 3 次元畳み込みニューラルネットワークの fine-tuning を行う際に利用する学習済みの畳み込みニューラルネットワークの比較を行った. 図 1 に示すように、異なるタスクのために学習されたニューラルネットワークを元として fine-tuning を行うことで、動画判別タスクにおいて、成績を向上させる fine-tuning の手法を検討する. 第 2 章の第 1 節では, 今回の検証に使用した畳み込み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, 3 次元畳み込みニューラルネットワークの fine-tuning について考察を行う.

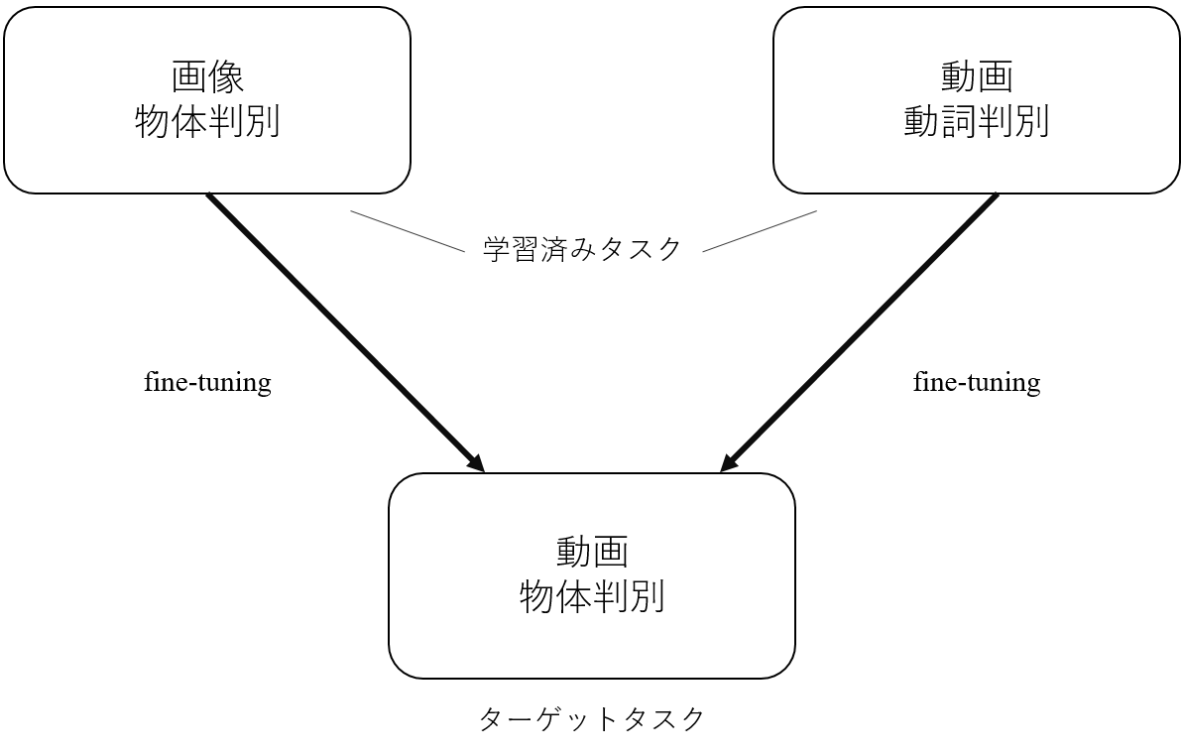


**図 1 本研究における比較の概念図.** 本研究においては, 動画中の物体判別タスクを学習するために, 複数の学習済み畳み込みニューラルネットワークを元に fine-tuning を行い, その結果を比較した. その際に, タスクは同じ物体判別であるが, 画像を判別するために学習されたニューラルネットワークと, 動画を対象とすることは共通しているが, 動詞判別タスクで学習を行ったニューラルネットワークを比較し, fine-tuning において対象とするデータの類似性とタスクの類似性が与える影響を検証した.

# 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, 訓練済み畳み込みニューラルネットワークを fine-tuning し, 物体判別タスクの検証を行った.

### 2.1.1 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, ほか, 2009]を用いた 1000 クラス判別タスクで pre-train された ResNets [He, Zhang, Ren, Sun, 2015] を用いた. 本検証においては, 50 層の ResNets を用いた.

### 2.1.2 3 次元畳み込みニューラルネットワーク

本研究では 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-train 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時空間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-train したネットワークを用いた.

#### 2.1.2.1 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, 2 次元画像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時空間畳み込みニューラルネットワークである. 拡張は, 時空間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. 時空間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, 時空間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替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限をみたすように行う.

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, 時空間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の重みを時間方向に N 回繰り返した後に, 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, 時空間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, 後者を中心化拡張と呼ぶ.

本検証においては, 画像中の物体判別に pre-train された時空間畳み込みニューラルネットワークとして, ImageNet で pre-train された ResNets50 をそれぞれ, 平均化拡張, 中心化拡張によって時空間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 動詞判別時空間畳み込みニューラルネットワーク

時空間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-train 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-train 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, ほか, 2018] から抽出した 1250 件の動画データ及び, 動画に対応する物体カテゴリラベルを使用した. 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カテゴリ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カテゴリラベルを利用するため, Moments In Time データセットから訓練, バリデーション用のデータとして 1200 件, テスト用に 50 件のデータを元としてデータセットを作成した. 訓練, バリデーション用データは 150 のアクションカテゴリから 8 件ずつ, テスト用データはその内 50 のカテゴリから 1 件ずつのデータを使用した.これらの動画に対する物体カテゴリのラベリングを, 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 の物体カテゴリ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, この内出現頻度上位 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 画素のサイズの物であった. 本検証においては, 90 フレームから均等に 32 フレームを抽出して用いた.

## 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て入力を行った. 入力は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 バッチとして入力を行った. 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 x 256 の解像度の画像から 224 x 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 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 を用いた正則化を行った.

### 2.3.2 時空間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, 1 動画を 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 x 256 の解像度の動画から 224 x 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

物体判別タスクの成績は以下の方法で検証した.

### 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, 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 x 256 の解像度の画像及び動画から中央の 224 x 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, 20 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 AUC (Area Under Curve) を用いて行った. 20 ラベルそれぞれについて, テストセット全体の予測値と真のラベルを用いてラベル毎の AUC を算出した.

# 3 章 結果

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

## 3.1 学習曲線

畳み込みニューラルネットワークの学習時の損失の比較を行った.

図 x は, ニューラルネットワークの学習中の損失の比較である. 損失は最終層の各カテゴリ毎の予測における交差エントロピー誤差を各カテゴリ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.

## 3.2 判別結果

マルチラベル判別問題の結果の比較を行った.

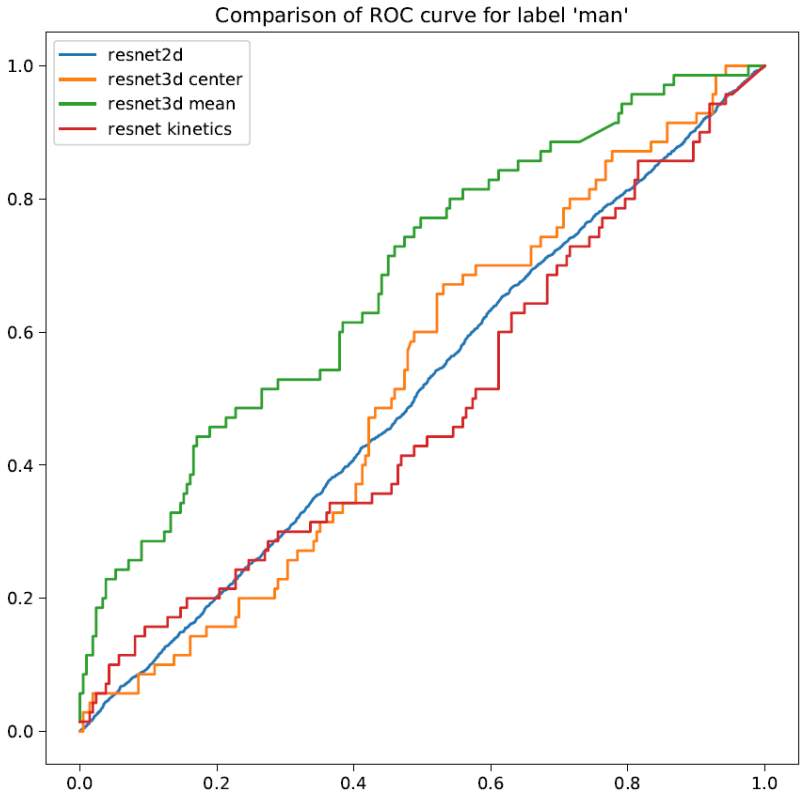


**図 2 カテゴリ判別の ROC 曲線.** 各畳み込みニューラルネットワークを元に fine-tuning したネットワークによる “man” ラベルの二値予測に対する ROC 曲線. 横軸が偽陽性率, 縦軸が陽性率を表す.

図 x は, いくつかのカテゴリに対する予測の 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時空間畳み込みニューラルネットワーク, 動詞判別時空間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時空間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.

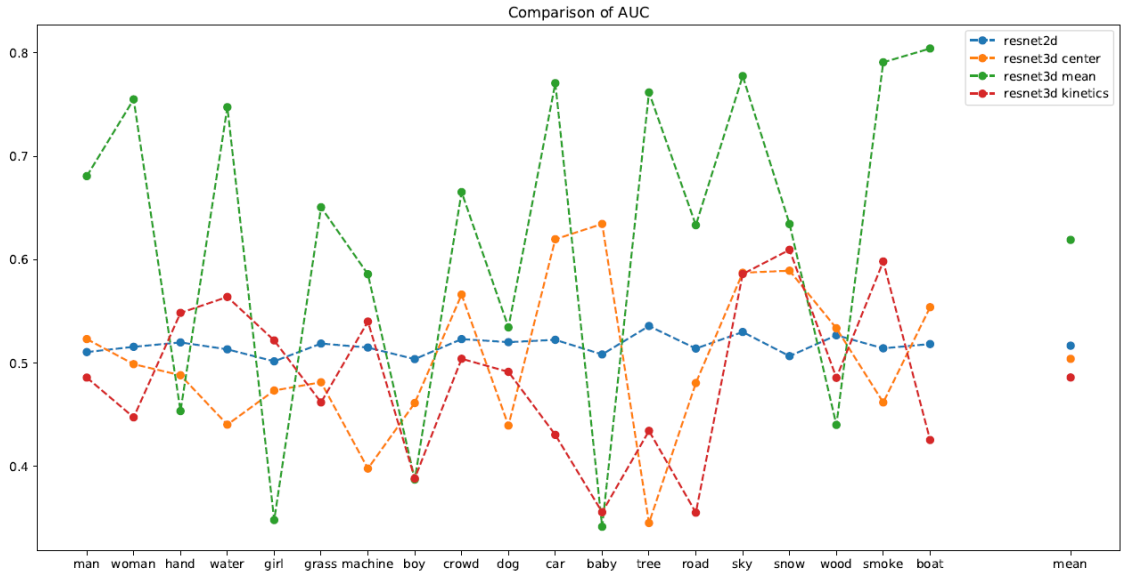


**図 3 カテゴリ毎の AUC 比較.** カテゴリ毎の AUC をそれぞれの畳み込みニューラルネットワーク毎に比較している. 縦軸は AUC の値. カテゴリのラベルは左からデータセット中での出現頻度が高いものから順に並んでいる. 一番右側にある mean ラベルは, 各畳み込みニューラルネットワークの AUC のカテゴリ平均を示している.

また, 図 x 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り判別結果では, 各カテゴリの AUC はおおよそ 0.5 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時空間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時空間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. 動詞判別時空間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.

# 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. 以下では, まず, それぞの畳み込みニューラルネットワークを用いて fine-tuning お行った際の性質について考察を行う. その後に, 4 つの畳み込みニューラルネットワークの比較から時空間畳み込みニューラルネットワークにおける fine-tuning の特性についての考察を行う.

fine-tuning をベースにした動画中の物体判別タスクにおいては, 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学習に成功したと言える. 図 x で示された通り, 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 x 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, 時空間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.

時空間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時空間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, 時空間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時空間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, 時空間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時空間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時空間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.

## 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

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

HeKaiming, ZhangXiangyu, RenShaoqing, SunJian. (2015 年 12 月 10 日). Deep Residual Learning for Image Recognition.

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

MonfortMathew, ZhouBolei, BargalSarah Adel, AndonianAlex, YanTom, RamakrishnanKandan, . . . OlivaAude. (2018). Moments in Time Dataset: one million videos for event understanding.

TranDu, BourdevLubomir, FergusRob, TorresaniLorenzo, PaluriManohar. (2015). Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision.