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 3~~10~~日 提出

書式変更: フォント : 14 pt

書式変更: フォント : 14 pt

目次

書式変更

要旨

Fine-tuning とはある課題のためにすでに訓練されたニューラルネットワークを元に, 別の課題のために再訓練することを指す. 一般にニューラルネットワークの学習には大規模なデータが必要となるが, fine-tuning では学習済みのニューラルネットワークを元とすることで新たな課題に対し, 少量の訓練データ量で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判別するタスクにおいては, 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 は複数の方法が提案されており, どの方法が優れているか確立された見解は得られていない. そこで本研究では動画中の物体判別タスクにおいて, 先行研究で提案されている複数の fine-tuning の方法を試し, 精度を比較した. 検証の結果, 静止画中の物体判別タスク用に学習されたネットワークを動画に適したアーキテクチャに拡張したネットワークを元に fine-tuning を行った場合には動画中の物体判別タスクを学習でき, それ以外の訓練済みニューラルネットワークを元に fine-tuning を行った場合にはチャンスレベルと同等の精度となることが分かった. 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. ~~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~~

コメントの追加 [間島慶1]: 比較した, と述べているのだから, 特定の場合で学習できたことだけでなく, 比較の結果を述べるべき.

~~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, 多様なタスク課題において革新的な性能の向上をもたらしてきた. ニューラルネットワークは大量のデータの学習により~~によりサ~~, 画像認識・音声認識・自然言語処理など様々なタスク課題において時にはヒトに勝る性能~~これまでにない~~を~~成果を~~出しており, 音声操作システムや顔認証システム, 自動翻訳などへの応用により我々の日常生活にも大きな影響を与えている~~きた~~. また, 学習に大量のデータが必要となる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学習~~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~~てることにより,~~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 (pre-training) , 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. Fine-tuning は~~その際に~~第一段階に用いる~~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できた~~している~~場合, ターゲットとするタスクの学習を比較的少量のデータで行うことができるため, 様々なタスクの学習において頻繁に用いられている~~必要がある~~.

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 fine-tuning によって学習~~学習~~することにより, fine-tuning を用いない場合よりも判別成績が向上することが~~が可能であることが~~示されている (Carreira & Zisserman, 2017) . また, 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を動画を扱うニューラルネットワークに拡張する手法として, ニューラルネットワークの平均化拡張 (Carreira & Zisserman, 2017) や中心化拡張 (Girdhar et al., 2018) が提案されており, これらの手法を用いて訓練済みの静止画用のニューラルネットワークを元に動画用のニューラルネットワークを fine-tuning できるという報告も上がっている (Carreira & Zisserman, 2017; Hara et al., 2017). しかし, 複数の候補がある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~~から, どの手法を用いるべきなのかについては共通の見解は生まれていない. ~~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, どのような ~~うまうな~~ 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~~訓練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 によって~~を行い~~動画中の物体判別タスクの学習を行った. 検証に際しては, 訓練済みニューラルネットワークとして, 表 1 に示すように 1) 静止画中の物体判別タスクを学習した静止画用のニューラルネットワーク, 2) 静止画中の物体判別タスクを学習した静止画用のニューラルネットワークを平均化拡張を用いて動画用のニューラルネットワークに拡張したネットワーク, 3) 静止画中の物体判別タスクを学習した静止画用のニューラルネットワークを中心化拡張を用いて動画用のニューラルネットワークに拡張したネットワーク, 4) 動画中の動詞判別タスクを学習した動画用ニューラルネットワークの 4 つを用いた. これらの訓練済みモデルを元に fine-tuning したモデルによる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.

訓練済みモデル	ネットワークアーキテクチャ	学習したタスク
2 次元 CNN	2 次元畳み込み (静止画用) ~~(静止画用)~~	静止画中の 物体判別タスク
中心化拡張 CNN	3 次元畳み込み (動画用) ~~(~~平均化拡張 (Carreira & Zisserman, 2017)	静止画中の 物体判別タスク
平均化拡張 CNN	3 次元畳み込み (動画用) 中心化拡張 (Girdhar et al,. 2018)	静止画中の 物体判別タスク
動詞判別 CNN	3 次元畳み込み (動画用)~~(動画用)~~	動画中の 動詞判別タスク

表 1. 検証に用いた訓練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訓練済みのニューラルネットワークの一覧. 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の異なる訓練済みモデルを利用した.

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 節で比較検討の方法論について述べた後, 第 3, 4 章では, 比較および検証を行い, ニューラルネットワークの fine-tuning について考察を行う.

書式変更: 両端揃え, インデント : 左 0.57 字, 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
表の書式変更
書式変更: インデント : 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: 両端揃え, インデント : 左 0.57 字, 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: インデント : 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: インデント : 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: 両端揃え, インデント : 左 0.57 字, 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: 両端揃え, インデント : 左 0.57 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: インデント : 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt, 次の段落と分離しない
書式変更: 両端揃え, インデント : 左 0.57 字, 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: インデント : 最初の行 : 0 字, 間隔 段落前 : 0 pt, 段落後 : 0 pt
書式変更: フォント : 太字, 斜体 (なし), (言語 1) 日本語
書式変更: インデント : 最初の行 : 0 字, 間隔 段落前 : 自動
書式変更: フォント : 太字, 斜体 (なし), (言語 1) 日本語
書式変更: フォント : 太字, 斜体 (なし), (言語 1) 日本語
書式変更: フォント : 太字, 斜体 (なし), (言語 1) 日本語
書式変更: 本文, インデント : 最初の行 : 0 mm

~~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~~

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 (CNN) を使用した. CNN ~~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, 本研究で対象とする静止画中の物体判別や動画中の動詞・物体判別タスクにおいて一般的に用いられている~~サ, 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~~ CNN 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~~ CNN を使用して検証を行った.

本検証では, 複数の訓練済み CNN ~~畳み込みニューラルネットワーク~~ を 元に fine-tuning することにより~~七~~, 動画中の物体判別タスクを学習し, その成績を比較することで 動画中の物体判別タスクの fine-tuning に用いるべき訓練済み CNN を検証した~~スクの検証を行った~~.~~-~~ 比較に用いた訓練済み CNN は表 1 で示しているように 2 次元 CNN, 平均化拡張 CNN, 中心化拡張 CNN, 動詞判別 CNN の 4 種類である ~~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.~~

2.1.1 2 次元 CNN~~2 畳み込みニューラルネットワークの拡張~~

2 次元 CNN として, ImageNet (Deng et al., 2009) を用いた静止画中の 1000 クラス物体判別タスクを学習した ResNets (He et al., 2016) を使用した. ResNets は, residual block という構造を複数重ねた CNN であり, 静止画認識の分野において飛躍的な成果を出し広く活用されている CNN である.

2.1.2 平均化拡張 CNN

~~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~~ CNN で, 静止画中の物体判別タスクを行うものの一つとして平均化拡張 CNN を用いた~~として~~. 本検証では~~,~~訓練済みの 2 次元 CNN を 3 次元に拡張することにより作られる ~~-~~I3D ネットワーク~~-~~ (Carreira & Zisserman, 2017) を用いることにより, 前述の静止画中の物体判別を行う 2 次元 CNN を拡張し, 静止画中の物体判別を行う 3 次元 CNN を作成した.~~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~~ CNN の作成と, 2 次元~~畳み込みニューラルネットワーク~~ CNN からの~~学習済み~~訓練済みの重みの転移によって作成される~~行われる~~. 3 次元~~畳み込みニューラルネットワーク~~ CNN は,~~ 畳み込みニューラルネットワーク~~ CNN 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~~ CNN 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~~元の 2 次元~~畳み込みニューラルネットワーク~~ CNN 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みたすように行う.

本検証では, 3 次元 CNN に拡張した ResNets50 に, この平均化拡張を用いて前述の静止画中の物体判別タスクを学習した 2 次元 CNN の重みを転移することによって初期化を行った CNN を平均化拡張 CNN と呼ぶ..

2.1.3 中心化拡張 CNN

中心化拡張 CNN も平均化拡張 CNN と同様に静止画中の物体判別タスクを行う 3 次元 CNN である. 平均化拡張 CNN と同様に 2 次元 CNN を 3 次元 CNN に拡張することで作成されるが, 2 次元の CNN から重みを転移する際の手法に中心化拡張を用いる. 中心化拡張は, 3 次元 CNN の畳み込み層の重みをすべて 0 で初期化した後に, 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 CNN の畳み

込み層の重みを転移することによって初期化を行う I3D ネットワークの作成方法である (Girdhar et al,. 2018) . 本検証では, 3 次元 CNN に拡張した ResNets50 に, この中心化拡張を用いて. 前述の静止画中の物体判別タスクを学習した 2 次元 CNN の重みを転移することによって初期化を行った CNN を平均化拡張 CNN と呼ぶ..

3 次元~~畳み込みニューラルネットワーク~~ CNN 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 (Kay et al., 2017) を用いた動画中の~~の~~動詞判別タスクで pre-training された 3 次元~~ニューラルネットワーク~~ CNN 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~~,~~ 3 次元 CNN ネットワークの~~ア~~アーキテクチャは~~構造としては前述のも~~平均化拡張 CNN や中心化拡張 CNN ~~の~~と同様の ResNets50 を 3 次元に拡張~~を使用~~したものを利用した~~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~~ CNN 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~~に物体ラベル付けを行ったデータセットを用いた~~, 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 データセットのラベリングおよび前処理~~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合計 ~~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 件の動画データを抽出し~~元としてデータセットを作成した.~~, それぞれに物体ラベルを付与することでデー

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,~~は, それぞれの動画中にを複数つける形で行った. ラベリングを行った結果, 193 種類の物体ラベルが動画に付与され~~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 件で, 1 動画あたりの平均のラベル数は 1.25 であった. この内, 70% を訓練用データ, 30% をテスト用データとして用いた. 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の学習

2 次元の~~畳み込みニューラルネットワーク~~ CNN は以下の方法で fine-tuning ~~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, 上下の反転をランダムに~~おこなった~~行った後, 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. これは一つの動画に対して複数のラベルを予測するタスクを 一つの動画に対して, 20 あるラベルがそれぞれ含まれているかをラベル毎に二値判別として予測するタスクと扱うと自然な損失関数となる. 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学習率を 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の学習

平均化拡張 CNN , 中心化拡張 CNN, 動詞判別 CNN の 3 つの 3 次元の~~畳み込みニューラルネットワーク~~ CNN は以下の方法で fine-tuning ~~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, 入力時にランダムに左右, 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,~~ 損失関数, 最適化手法, ~~学習時の~~ハイパーパラメータ~~条件~~は上述の 2 次元~~畳み込みニューラルネットワーク~~ CNN と同様のものを用いた.

2.~~4~~3 検証

2.~~4~~3.1 評価方法

~~畳み込みニューラルネットワーク~~ Fine-tuning 手法による動画中の物体判別タスク CNN に ~~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~~に行った~~用いて行った~~. ~~畳み込みニューラルネットワーク~~ CNN への入力は 256 × 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~~ CNN の最終ルの予測値として評価を行った.

2.~~3~~4.2 評価指標

ニューラルネットワークによる予測の評価は, それぞれの物体ラベル~~が~~~~の物体が予測~~動画~~画像~~に含まれているかの二値判別問題として Area Under the Curve (AUC) を用いて行った. 20 ラベルそれぞれについて, テストデータ~~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.~~

Fine-tuning を行った~~する~~際の学習曲線の比較を行った. 損失は CNN の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. 図 1 は各 CNN の学習曲線の比較である. 全ての~~畳み込みニューラルネットワーク~~ CNN 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~~ CNN 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~~ CNN 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に~~は~~大きな差は見られなかった.

図 ~~2~~1 ~~畳み込みニューラルネットワーク~~ CNN の学習曲線 それぞれの CNN ~~畳込みニューラルネットワーク~~の訓練時の~~学習時~~損失~~ロス~~の比較~~を行った~~. 左上が 2 次元~~畳み込みニューラルネットワーク~~ CNN , 右上が中心化拡張 CNN ~~ニューラルネットワーク~~, 左下が平均化拡張 CNN ~~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~~ CNN 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.

動画中の物体判別問題の結果の比較を行った. 図 ~~3~~2 は, データセットにおける最頻~~いくつかの~~ラベルである “man” 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~~ CNN , 動詞判別 ~~3 次元畳み込みニューラルネットワーク~~ CNN, 2 次元~~畳み込みニューラルネットワーク~~ CNN においては ROC 曲線はチャンスレベルのものと同等ので, 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~~ CNN

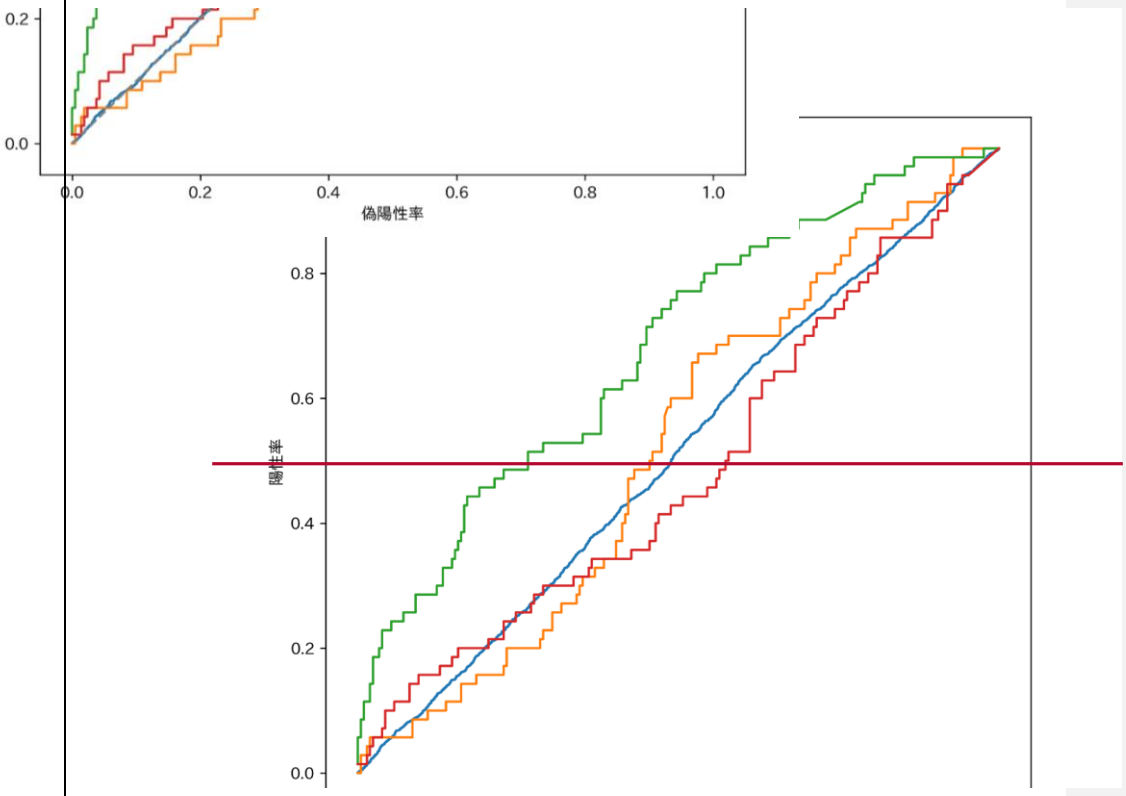

図 ~~3~~2 -ラベル判別の ROC 曲線. 各~~畳み込みニューラルネットワーク~~ CNN を元に fine-ラベルの二値予測に対する ROC 曲線の比較を行った. 破線はチャンスレベルを示している.

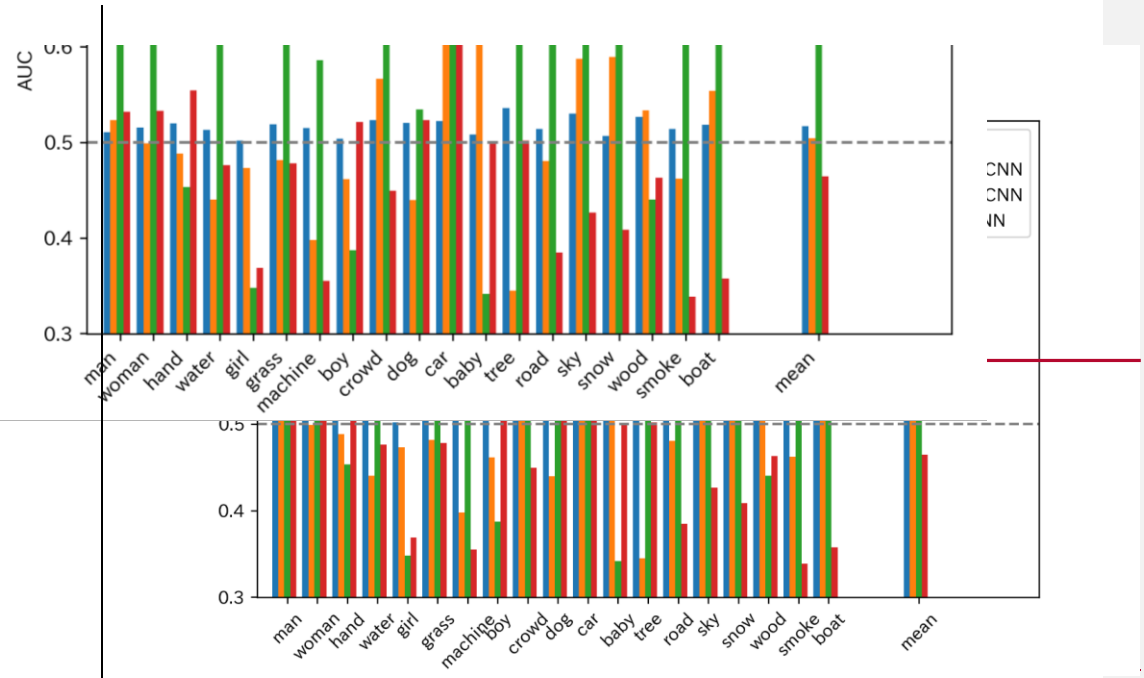

図 43 ラベル毎の AUC の比較. ラベル毎の AUC をそれぞれの~~畳み込みニューラルネットワーク~~ CNN 毎に比較している. 縦軸は AUC の値. ~~ラベルの~~ラベルは左からデータセット中での出現頻度が高いものから順に並んでいる. -一番右側にある mean ラベルは, 各~~畳み込みニューラルネットワーク~~ CNN の AUC のラベル平均を示している.

第 4 章 考察

本研究では, 複数の訓練済み~~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~~ CNN ~~の~~を元に動画中の物体判別タスクの fine-tuning を行い, ~~判別動画中の物体判別タスクの~~成績の比較を行った. ~~結果,~~ その結果, 今回比較, fine-tuning に用いた~~を行った~~ 4 つの~~畳み込みニューラルネットワーク~~ CNN 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~~ CNN のみが他のものと比べ高い成績を示すことが分かった.

2 次元~~の畳み込みニューラルネットワーク~~ CNN から fine-tuning を行った場合ストデータに対する判別誤差~~成績~~を大きく下回る過学習が起きることが分かった. 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~~ CNN 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学習の初期段階で一定となることが明らかになった.

以上のような比較から, 動画中の物体判別タスクの ~~3 次元畳み込みニューラルネ~~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~~ 3 次元~~畳み込みニューラルネットワーク~~ CNN が挙げられるが, 本検証に用いた~~よう~~場合には 2 次元の~~畳み込みニューラルネットワーク~~ CNN は過学習に陥る傾向があ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~~中の物体判別タスク~~~~を用いたタスクに~~の fine-tuning における前述の~~した~~問題を緩和する方法として, 物体判別という点で共通する静止画中の物体判別~~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~~ CNN 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~~ CNN に拡張した~~ネットワーク~~平均化拡張 CNN を元として fine-tuning を行う方法があることが明らかになった~~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~~ CNN の~~学習済み~~訓

練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における ~~学習のための動画の~~ fine-tuning の特性を調査するために, 複数の訓練済み ~~異なる学習済み畳み込みニューラルネットワーク~~ CNN を用いて fine-tuning を行い, 動画中の物体判別タスクにおける判別成績 ~~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, 静止画中 ~~同様~~の物体 ~~画像~~判別タスクで ~~学習済み~~訓練済みの 2 次元 ~~畳み込みニューラルネットワーク~~ CNN 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~~ CNN に拡張した訓練済み CNN ~~ネットワーク~~を元として fine-tuning を行った際に判別成績が高く, その他の訓練済み CNN を用いた場合にはチャンスレベルと同等の判別成績となることが分かった ~~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~~. その原因の仮説として, 動画に対して 2 次元 CNN を用いることで過学習が起きやすくなること, 限られた量のデータで fine-tuning を行う際には fine-tuning 前後での CNN の重みの変化量が少ない方が学習が成功しやすい可能性があることが明らかになった. また, 平均化拡張 CNN で一定学習が進んだことは, 本検証のように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.

コメントの追加 [間島慶2]: 比較した, と述べているのだから, 特定の場合で学習できたことだけでなく, 比較の結果を述べるべき.

~~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.~~

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. また, ATR 脳情報研究所 の皆様には, 論文の推敲など執筆にあたりご指導, サポートをしていただき感謝いたします. 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., Courville, A., ... & 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, J., Wei Dong, W., Socher, R., Li Jia Li, L., Kai Li, K., & Li Fei-Fei, L. (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 Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (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 Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., & Oliva, A. (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)~~.

Hara, K., Kataoka, H., & Satoh, Y. (2017). Learning spatio-temporal features with 3D residual networks for action recognition. In *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition* Vol. 2. No. 3..