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 31日 提出

目次

[要旨 1](#_Toc536624115)

[第1章 序論 2](#_Toc536624116)

[第2章 方法 5](#_Toc536624117)

[2.1 畳み込みニューラルネットワーク 5](#_Toc536624118)

[2.1.1 2次元 CNN 5](#_Toc536624119)

[2.1.2 平均化拡張 CNN 5](#_Toc536624120)

[2.1.3 中心化拡張 CNN 6](#_Toc536624121)

[2.1.4. 動詞判別 CNN 6](#_Toc536624122)

[2.2 データセット 7](#_Toc536624123)

[2.2.1 Moments In Time データセット 7](#_Toc536624124)

[2.2.2 データセットのラベリングおよび前処理 7](#_Toc536624125)

[2.3 物体判別学習 7](#_Toc536624126)

[2.3.1 2次元畳み込みニューラルネットワークの学習 7](#_Toc536624127)

[2.3.2 3次元畳み込みニューラルネットワークの学習 8](#_Toc536624128)

[2.4 検証 8](#_Toc536624129)

[2.4.1 評価方法 8](#_Toc536624130)

[2.4.2 評価指標 9](#_Toc536624131)

[3章 結果 10](#_Toc536624132)

[3.1 学習曲線 10](#_Toc536624133)

[3.2 判別成績 12](#_Toc536624134)

[第4章 考察 14](#_Toc536624135)

[第5章 結論 16](#_Toc536624136)

[謝辞 17](#_Toc536624137)

[参考文献 18](#_Toc536624138)

**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.

**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, 画像認識・音声認識・自然言語処理など様々なタスクにおいて時にはヒトに勝る性能を出しており, 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. 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 は第一段階に用いるタスクの学習を通して, ニューラルネットワークが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. 一例として, 静止画を扱うタスク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有用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fine-tuningの際の元のモデルとして使用されている. このように, 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 fine-tuning して静止画を扱うタスクを学習する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fine-tuning によって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.

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 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訓練済みニューラルネットワークを用意し, fine-tuning によって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.

**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, 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 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. 本研究では静止画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 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 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 を 元にfine-tuning することにより, 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. 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. 各訓練済み CNN について以下に詳述する.

**2.1.1 2次元 CNN**

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**

3 次元 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. 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.

I3Dネットワークは 3次元 CNN の作成と, 2次元 CNN からの訓練済みの重みの転移によって作成される. 3次元 CNN は, 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 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を入力した時の出力が, 元の2 次元 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.平均化拡張はこのような制約をみたす CNN の拡張方法の一つであり.変換先の3次元 CNN の畳み込み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保ったまま時間方向に*N*回重ねた後に, 重みの値を1/*N*倍することにより 3次元CNN の重みの初期化を行う手法である (Carreira & Zisserman, 2017).

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 の畳み込みそう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..

**2.1.4. 動詞判別 CNN**

3次元 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) を用いた動画中の動詞判別タスクで pre-training された3次元 CNN を用いた. 3次元 CNN のアーキテクチャは平均化拡張 CNNや中心化拡張 CNN と同様のResNets50 を3次元に拡張したものを利用した.

**2.2 データセット**

**2.2.1 Moments In Time データセット**

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 件の動画データに物体ラベル付けを行ったデータセットを用いた. 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**

本研究では, Moments In Timeデータセットから訓練, テスト用のデータとして1200件の動画データを抽出し, 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. ラベリングは, 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 種類の物体ラベルが動画に付与され, 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 CNN は以下の方法で fine-tuning 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,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ラベルのそれぞれが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である. 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 CNN は以下の方法で fine-tuning 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, ハイパーパラメータは上述の2 次元 CNN と同様のものを用いた.

**2.4 検証**

**2.4.1 評価方法**

Fine-tuning 手法による動画中の物体判別タスクの成績の比較は, fine-tuning 後のCNNに作成したデータセットのテストデータを入力した際のラベルの予測値を元に行った. CNN 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. CNN 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.

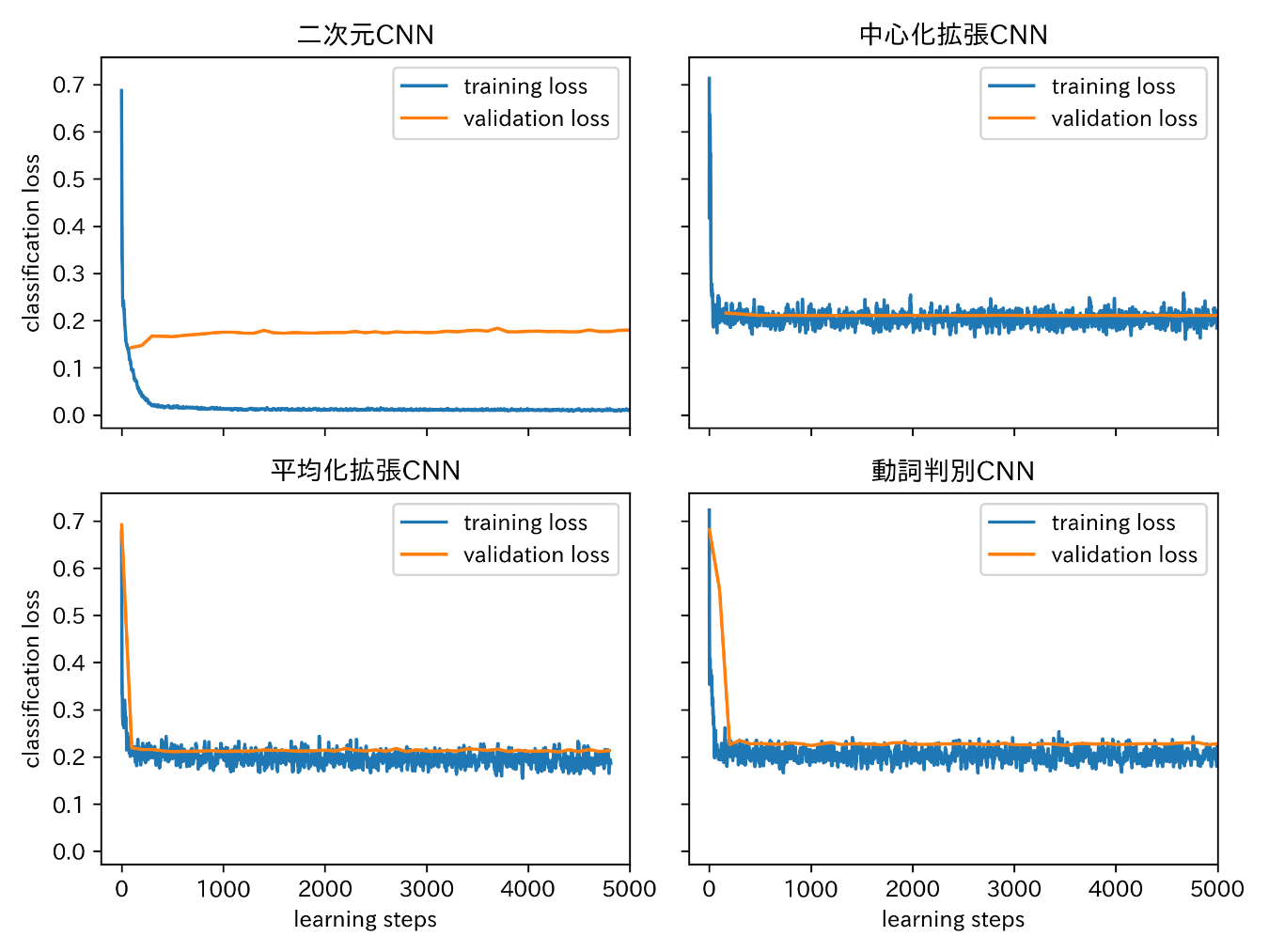
**2.4.2 評価指標**

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**

**3.1 学習曲線**

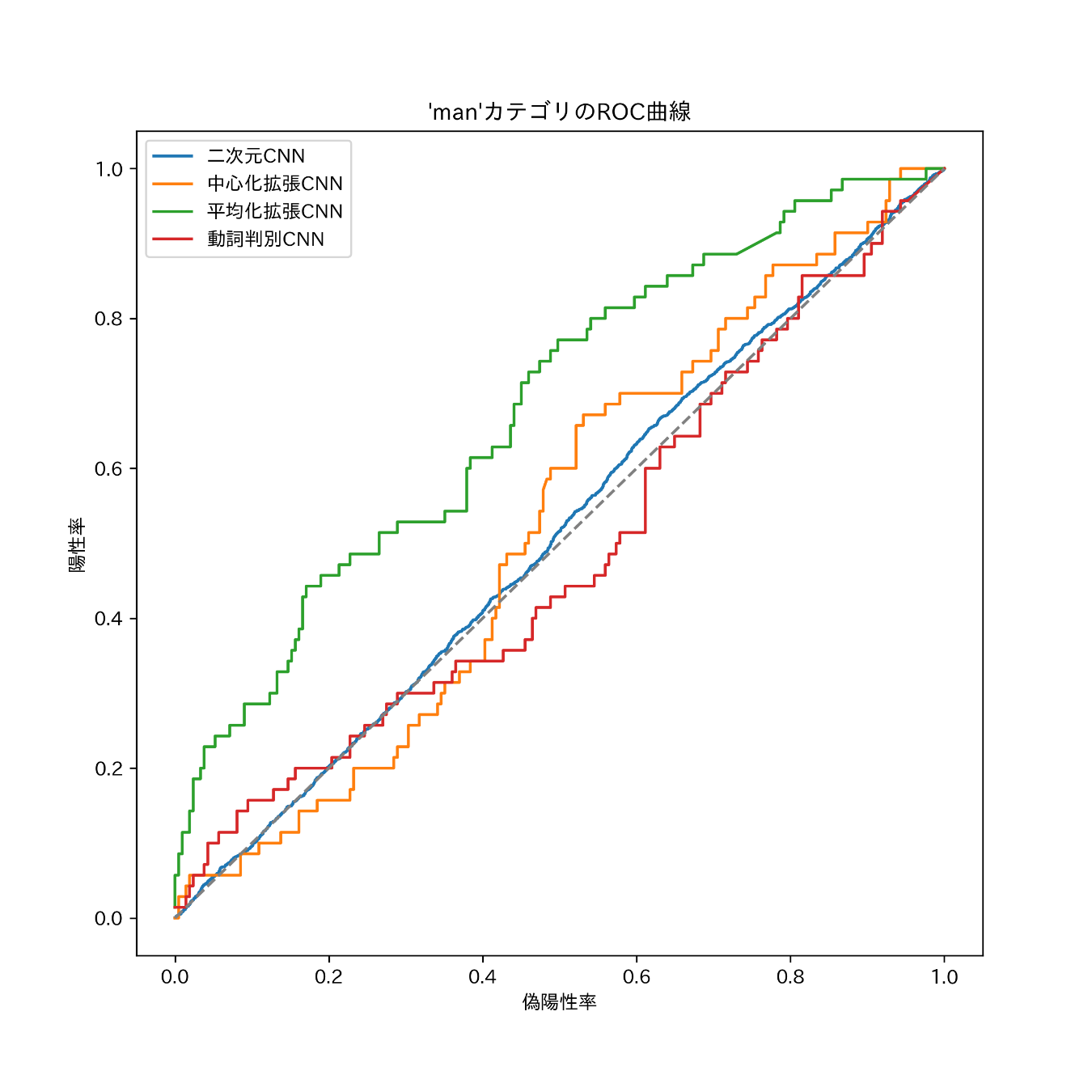
Fine-tuningを行った際の学習曲線の比較を行った. 損失は CNN の最終層の各ラベル毎の予測における交差エントロピー誤差を各ラベルにおいて平均することで求めた. 図1 が学習曲線である. 全ての CNN において学習初期に急激に損失が減少した後に学習が収束した. 2 次元 CNN においてのみ, 訓練データでの損失とテストデータでの損失に大きな差が見られ, それ以外の CNN においては, 訓練データとテストデータにおいて損失の値に大きな差は見られなかった.



**図 1 CNN の学習曲線** それぞれのCNNの訓練時の学習時損失の比較. 左上が2次元 CNN , 右上が中心化拡張 CNN, 左下が平均化拡張 CNN, 右下が動詞判別 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.

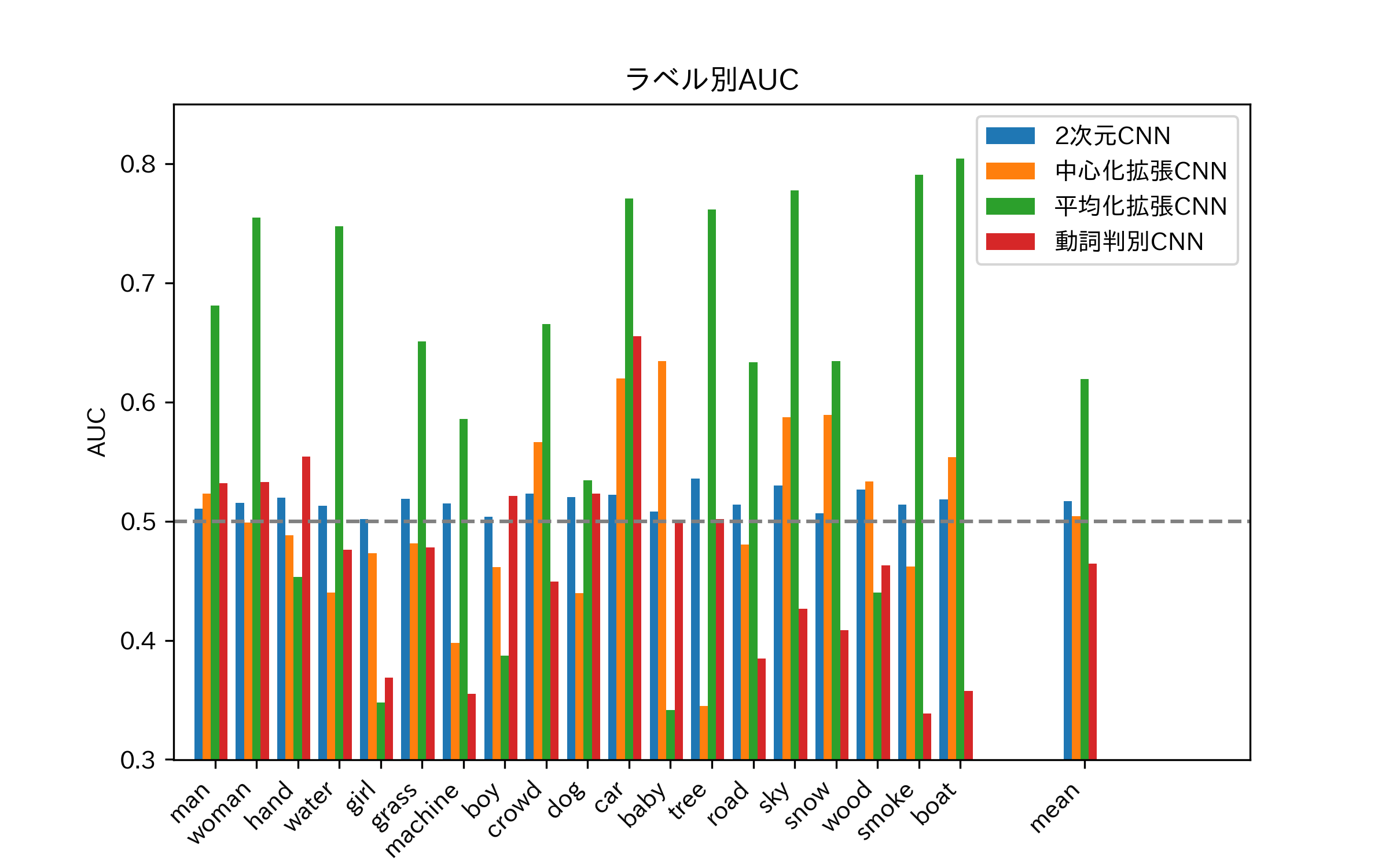
**3.2 判別成績**

動画中の物体判別問題の結果の比較を行った. 図2 は, データセットにおける最頻ラベルである “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CNN , 動詞判別 CNN, 2 次元 CNN 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 CNN は, 他の CNN よりも判別成績がよいことが分かった.



**図 2 ラベル判別のROC曲線.** 各 CNN 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曲線の比較. 破線はチャンスレベルを示している.

また, 図3 は CNN 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った. また, 中心化拡張 CNN においては, 2次元 CNN 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った. 平均化拡張 CNN 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, 学習が一定の成功を収めていることが分かった. 動詞判別 CNN 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 3 ラベル毎のAUC比較.** ラベル毎のAUCをそれぞれの CNN 毎に比較している. 縦軸はAUCの値. 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 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 CNN を元に動画中の物体判別タスクの fine-tuningを行い判別成績の比較を行った. その結果, 今回比較を行った4つの CNN の中で平均化拡張 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.また, その他の3つの CNN 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, 動画中の物体判別タスクの学習は成功しなかった. しかし, 図1 の学習曲線から, 学習が成功しなかった原因が2 次元 CNN と3次元 CNN において異なることが明らかになった.

2 次元 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 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.3次元 CNN の中で平均化拡張 CNN のみが物体判別タスクにおいて成績が高かった原因は, 平均化拡張においては初期の CNN 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からである可能性がある. 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 CNN を用いた場合に関しても, 物体判別と異なる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が考えられる. 本検証においては, 訓練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, 訓練済み CNN の重みの分析を行う必要がある.

以上のような比較から, 動画中の物体判別タスクの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 CNN としては2 次元の CNN と3次元 CNN 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 CNN は過学習に陥る傾向がある. 一方, 3次元 CNN 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, 動画中の物体判別タスクの fine-tuning おける前述の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タスクで訓練された2 次元 CNN を平均化拡張によって3次元 CNN に拡張した平均化拡張 CNNを元としてfine-tuningを行う方法があることが分かった. 今後の課題としては, fine-tuningを行った後の CNN の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 fine-tuning の特性を調査するために, 複数の訓練済み 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, 静止画の画像判別タスクで訓練済みの2 次元 CNN を平均化拡張を用いて3次元 CNN に拡張した訓練済み CNN 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 を用いた場合には学習は進まず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. 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ことを提案した. また, 平均化拡張 CNN では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.

**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．ATR 脳情報研究所の田中美里研究員には, 論文の推敲にお世話になり感謝いたします. 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., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, 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*, 770–778.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & 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., 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., 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*, 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*, 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*, 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, 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..