内容

[第1章 序論 2](#_Toc533583746)

[第2章 方法 3](#_Toc533583747)

[2.1 畳み込みニューラルネットワーク 3](#_Toc533583748)

[2.1.1 二次元畳み込みニューラルネットワーク 3](#_Toc533583749)

[2.1.2 三次元畳み込みニューラルネットワーク 3](#_Toc533583750)

[2.2 データセット 5](#_Toc533583751)

[2.2.1 Moments In Time データセット 5](#_Toc533583752)

[2.2.2 データセットの抽出 5](#_Toc533583753)

[2.3.1二次元畳み込みニューラルネットワーク 5](#_Toc533583754)

[2.3.2 時空間畳み込みニューラルネットワーク 6](#_Toc533583755)

[2.3 検証 6](#_Toc533583756)

[2.3.1 評価方法 6](#_Toc533583757)

[2.3.2 評価指標 7](#_Toc533583758)

[3章 結果 7](#_Toc533583759)

[3.1 学習曲線 7](#_Toc533583760)

[3.2 判別結果 7](#_Toc533583761)

# 第1章 序論

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の有用性は二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. 一方で, 三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. 特に, 動画中の物体識別タスクにおいて, 二次元学習の動画識別タスクで学習されたモデルと, 3次元の動画中の動詞判別タスクで学習された畳み込みニューラルネットワークを比較することにより, タスクの特性から抽出された特徴と, 画像に時間方向が加わった動画というデータの特性から抽出された特徴を比較することにより3次元畳み込みニューラルネットワークを用いた畳み込みニューラルネットワークにおけるfine-tuningの特性について検証を行う. 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, 三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。

# 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 二次元畳み込みニューラルネットワーク

二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 (引用) を用いた1000クラス判別タスクでpre-trainされた ResNets (引用) を用いた. 本検証においては, 50層のResNetsを用いた.

### 2.1.2 三次元畳み込みニューラルネットワーク

1. 本研究では三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された二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三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 (引用) ネットワークを用いた.

I3Dネットワークは, 訓練済みの二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二次元の同じ画像を繰り替し, 作成された動きがない動画 (boring-video) を入力した時の出力が, もとの二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のみを対応にする二次元畳み込みニューラルネットワークの重みを用いて初期化することによって転移を行った. 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

## 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 (引用) 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件のデータを抽出した. 抽出に際して訓練, 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件ずつのデータを抽出した. 抽出した動画に対する物体カテゴリのラベリングは, 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ラベルのみを用いた. 上位20のラベルがつけられた動画は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

### 2.3.1二次元畳み込みニューラルネットワーク

二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 (引用?) を, Momentumの値を0.9として使用した. 重みは4バッチ毎に勾配を蓄積し, その勾配を用いて更新した. 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 (引用？) 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

二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, 上述の二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二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. 二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。