2017年度　修士論文

位置変化に頑健なAttention付き Convolutional Neural Networkの提案

2018年1月30日　提出

指導教授

小林 哲則 教授

早稲田大学 基幹理工学部 情報理工学科
知覚情報システム研究室

5116F003-1

浅 野　秀 平

目 次

表 目 次

図 目 次

概要

本論文では，物体の位置変化に対しより頑健な画像処理を目的として Convolutional Neural Network(CNN) に attention の構造を取り入れる手法を提案する．

近年，ニューラルネットの一手法である CNN は，物体認識や背景分離といった幅広い画像処理のタスクに用いられている．CNN は，pooling と呼ばれる特徴マップ上の局所的な特徴をまとめ上げる処理を含み，画像処理において重要な物体の位置ずれに対する頑健性を有している．しかし，pooling による不変性は限定的であり，大域的な位置変化を含むデータを効率的に処理する事が出来ない．また，pooling を重ねると特徴マップの空間解像度が指数的に低下していき，物体がどこにあったかというレアイアウトに関する情報が失われてしまう．これらの性質は，物体の位置ずれを形状の固定された静的な受容野によって吸収しようとする為に生じる．

一方，人は広大な環境にあっても，高い空間解像度を維持しながら複雑な処理を高速に行っている．これは視覚的注意によって，環境の中から処理する領域を取捨選択する事が可能な為である．この仕組みでは，物体の像が視野の中央からずれていても，スポットライトのように注意が動く事で出力を一定に保つことが出来る．

そこで本研究では，この人の注意の仕組みを参考に，CNN に注意の構造を取り入れた Attentive Convolutional Neural Network(ACNN) を提案する．ACNN は通常の CNN の畳み込み層と全結合層の間に，空間的な attention を生成するネットワーク (attention network) を持つ．Attention を入力画像に応じて変化させる事で，全結合層へ入力する範囲を制御する事が可能となる．特徴マップからの部分的な特徴の読み出しは，attention と特徴マップの単純な行列積によって実現される．この行列積をニューラルネットの重みを掛ける操作として解釈すると，ACNN は入力データに応じてモデルの構造が変化するメタネットワーク [22] としての側面を持つ．学習には位置の情報といった追加の教師データを必要とせず，既存の CNN をそのまま ACNN に置き換える事が可能である．また，物体の配置に関する情報は attention network によって抽出され保存される．

検証では，平行移動を加えた手書き数字データセット MNIST に対する性能比較を通常の CNN と行った．ACNN は CNN の 1 割ほどのパラメータで，同等以上の識別性能を持つ

事を定量的に確認した．また，少量データを用いた検証により，ACNN が CNN よりも過学習を起こしにくい性質を持つ事を確認した．更に，モデルにとって未知のアフィン変換に対しても，ACNN は高い汎化性能を持つ事を確認した．

第1章　序論

1.1　研究背景と目的

机や椅子の置かれた教室や多くの人が行き交う交差点など，我々のいる実世界は複数の形状も大きさも異なる物体が多様な規則に沿って配置された，非常に複雑でバラエティに富んだ構造を持つ．一般に，このような実世界の画像を扱うシステムは，画像中の物体の位置ずれやスケールの変化に対して応答が不変である事が望ましい．例えば，歩行者検出のようなタスクでは，対象が画像内を自由に動き回っても，検出結果が保たれるようにシステムを設計しなければならない．

近年，Convolutional Neural Network (CNN) [8] は，物体認識 [9,11] や前景分離 [12]，画像生成 [13,14] といった幅広いタスクで用いられ優れた性能を示している．CNN は，図 1.1 のように，畳み込みを行うネットワークと全結合層を直列に並べた構造を持つ．畳み込みを行うネットワークは，更に局所的な特徴抽出を行う畳み込み層と，特徴を局所的にまとめ上げる pooling 層に分けられる．Pooling 層は，局所領域内の最大値や平均値といった代表値を出力とする事で，局所領域内の特徴の配置が変わっても同じ反応を保つことが出来る．しかし，その範囲は一般的に 2 × 2 程度の微小な大きさしか持たず，単層ではわずかな位置の変化に対してしか不変性を持たない．より大きな位置の変化を扱うには，畳み込みと pooling 層を積み重ねる必要がある [10]．しかし，入力画像に写った物体の複雑さに関係なく，単にその大きさに応じてネットワークの層を積み重ねるのは，計算資源や学習の収束速度の観点から効率的であるとは言えない．また，pooling を行うと空間解像度が指数的に落ちる為，画像の持つ空間的な情報 (物体の配置や傾き) は層を重ねるごとに失われてしまう．

YOLO [21] や SSD [21] といった最新の一般物体検出の手法では，複数のスケールとアスペクト比を持った anchor box を用いる事で物体の大きさや形状の違いに対処している．しかし，物体の位置の検出はあくまで画像全体を走査する事で実現されており，CNN と同様に不変性には限界がある．

一方，人の視覚は高い空間解像度を保ったまま，複雑な処理を同時かつ高速に行う事が

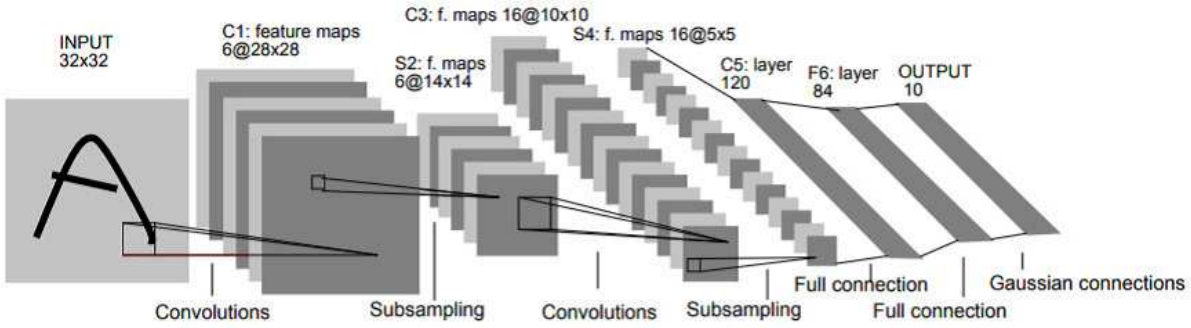

図 1.1: CNN のアーキテクチャ例 (LeNet-5)．[8] より引用．

可能である．この能力は，視覚的注意とよばれる有限な認知資源を空間に割り振る機構によって実現される．認知心理学の分野ではこの視覚的注意に関する数多くの知見が積み上げられており，そのメカニズムが明らかになりつつある．人の視覚的情報処理は，図 1.2 のように，視覚的注意を適用する前後で，異なった性質を持つ事が分かっている [1]．前注意的過程では，色や傾きといった単純特徴を視覚空間全体に渡って並列に計算する事が可能である．しかし，単純特徴を組み合わせて行う処理 (例えばリンゴを認識するには色と形状特徴を組み合わせる必要がある) を行う事は出来ない．次に集中的注意過程では，単純特徴を組み合わせた有機的な認知を行う事が可能である．しかし，処理容量に限りがあって，一度に狭い領域しか処理できず，注意による誘導を必要とする．この枠組みでは，特徴マップの空間解像度を高く保ちつつ，物体の位置の変化は注意が動くことで吸収する事ができる．本研究では，これらの人の認知の仕組みを参考に，CNN に注意の構造を加えた Attentive Convolutional Neural Network(ACNN) モデルを提案する．

1.2 提案手法の特徴

提案手法である ACNN は，CNN の畳み込み層と全結合層の間に空間的な attention を生成するネットワーク (attention network) を加えた構造を持ち，attention を変化させる事で全結合層へ入力する特徴マップの領域を制御する．Attention による特徴マップからの特徴の読み出しは，単純な行列積によって実現される．物体が大きく位置を変化させても，attention network がその位置を捉え，attention を変化させる事で全結合層への入力を一定に保つ．また，ACNN は特徴マップの一部のみを全結合層の入力として用いる為，モデルの持つパラメータ数は CNN と比較して大幅に小さく出来る．更に，attenton netowrk 内に

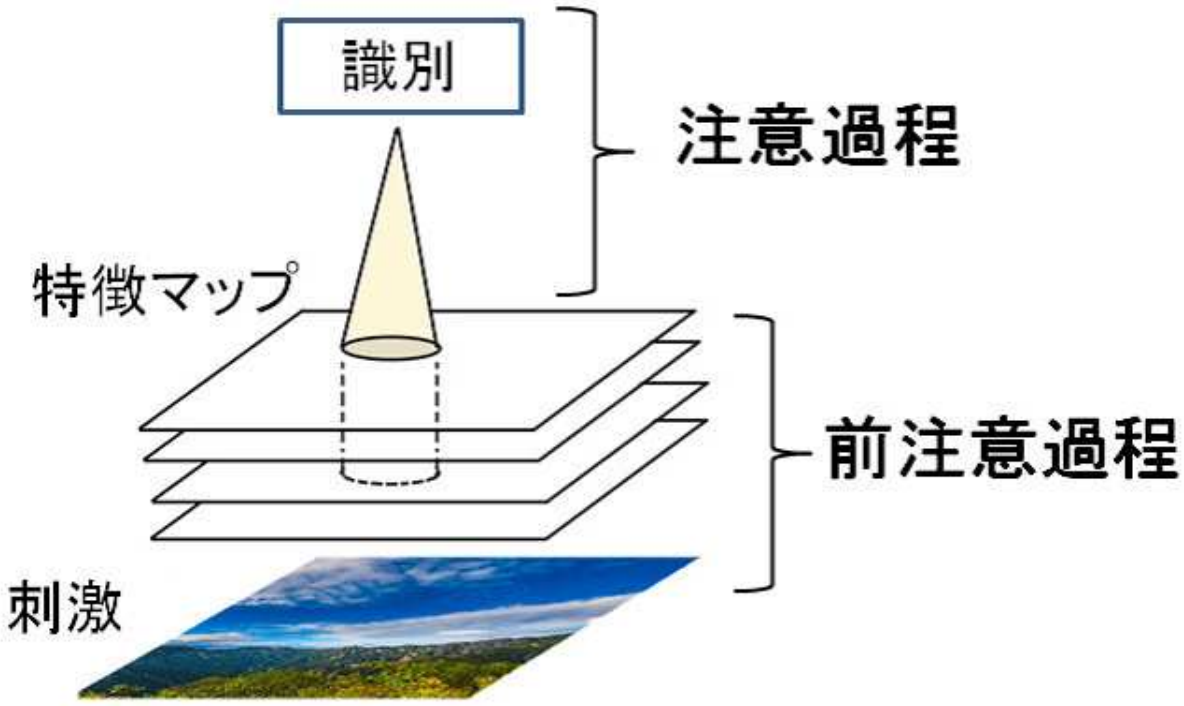

図 1.2: 人の視覚的注意の模式図.

は識別に必要な情報 (物体) の位置が保存されており，画像の再構成などに利用可能である.

第2章　関連研究

2.1　一般物体検出アルゴリズム

画像内の物体の位置とそのクラスを分けて処理するアルゴリズムは，その両方の推定を目的とする一般物体検出の分野において盛んに研究が行われてきた．一般物体検出を実現する方法として，画像内から複数の物体候補領域を抜き出し，各候補領域について物体のクラスを推定する方法が考えられる．Girshickらの提案したR-CNN [17]は，図2.1のように，候補領域の推定をSlective Search [16]と呼ばれる画素間の類似度に基づいたセグメンテーションによって行い，切り出した各候補領域に対して個別にCNNによる特徴抽出とSVMによるクラス分類を行った．以降，このRCNNをベースとしてアルゴリズムの高速化 [15,18]や候補領域の推定とクラスの推定を一つのニューラルネットによって学習する試み [19]が行われてきた．しかし，候補領域の提案に基づく物体検出は手続きが複雑で計算負荷が大きく，リアルタイムな処理や計算資源の限られた組み込みシステムへの応用が難しいという問題があった．

近年では，物体の位置の検出とクラスの推定をネットワークの中で並列に行う手法 [20,21]が提案されている．これらの手法では，畳み込みによって作成した特徴マップ上を，複数の形状と大きさを持った矩形 (default box)によって走査し，各矩形領域で位置とクラスのスコアリングを同時に行う事で高速化を実現している．

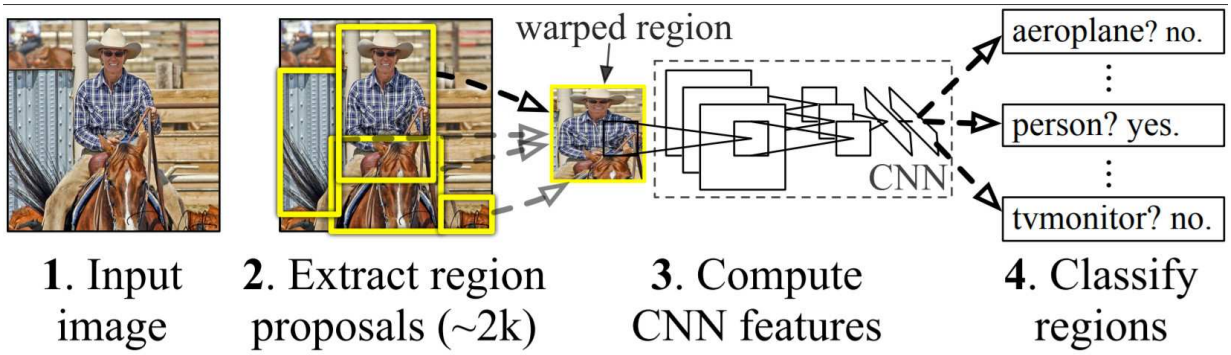

図 2.1: RCNNの処理過程．[17]より引用．

2.2 畳み込み層の受容野を変形させる試み

ある CNN の畳み込み層のユニットが影響を受けている入力画像の範囲を受容野 (receptive field) と呼ぶ．通常の CNN の場合では，この受容野は全て同じ大きさと形状を持ち，入力画像全体に等間隔に敷き詰められている．しかし，物体の形状や大きさは本来その種類や状態によって千差万別であり，それらを検出するのに最も適した受容野の形状もまた異なる．例えば歩行者の検出を行いたい場合には，受容野の形状は縦長であることが望ましい．そこで近年，畳み込み層の受容野の形状や位置をデータに合わせて最適化する研究が行われている．

Cheung らの研究 [3] では，人の視覚が周辺視と中心視で空間解像度が異なる事を参考に CNN の受容野の構造の最適化を行った，受容野の配置を 2 次元正規分布の集合として表現し，各正規分布の分散と平均 (位置) を最適化することで柔軟な受容野の配置を実現した．また，Jeon らの研究 [4] では，畳み込み層の各受容野に接続位置の変位量を直接パラメータとして持たせ，学習を通して最適化を行った．しかし，これらの研究では受容野の形状そのものを学習パラメータとして最適化しており，学習後はその形状や位置が変化する事はない．

次に Jaderberg らの研究 [6] では，図 2.2 のように，物体の位置や回転の検出を目的としたネットワーク (localisation network) を畳み込み層に追加し，localisation network から出力したアフィン変換のパラメータによって受容野の位置を変形した．しかし，変形はアフィン変換で表現可能なものに限られ，局所的に密なサンプリングを行う事はできなかった．Dai らの研究 [5] では，受容野毎に接続位置の変位量を直接決定するネットワークを新たに畳み込み層に付け加えた．任意の配置の受容野を表現可能になったが，各受容野の変位量を決定するネットワークが畳み込み層によって構成されており，ある受容野の変位量を同じ受容野を持つ別のネットワークが決定する構造となっている．よって，単層では受容野の外の状況に応じて受容野の変位量を決める事が出来ず，変形のモデリング能力を上げる為には層を積み重ねる必要がある．

これらの受容野の変位量を連続変数として持つ手法は，特徴マップから特徴を読み出す際に，変位量を連続変数から整数座標に床関数等を用いて離散化する必要がある．この離散化の操作は微分不能であり，誤差逆伝搬を用いてネットワークを学習する上で問題となる．そこで，適当なカーネルによる補間による微分可能化の操作を必要とするが，これは

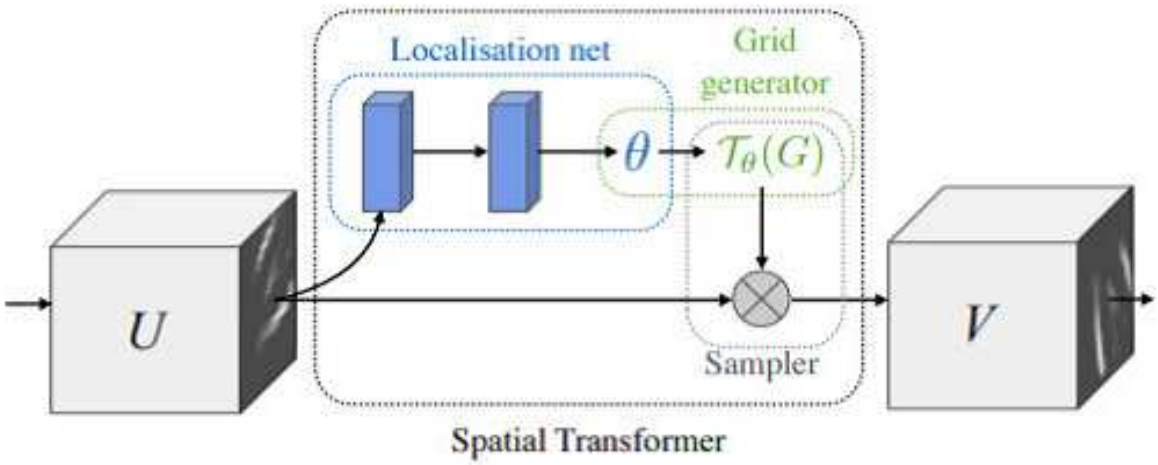

図 2.2: Spatial Transformer Network. [6] より引用.

高いオーバーヘッドとなる [4].

第3章　Attentive Convolutional Neural Network

本章では，CNN に注意の構造を取り入れた，Attentive Convolutional Neural Network (ACNN) モデルの構造について述べる．

3.1　Attentive Convolutional Neural Network

まず，ACNN の概要を図 3.1 に示す．ACNN は，CNN の畳み込みネットワークと全結合ネットワークの間に空間的な attention を生成するネットワーク (attention network) を持つ．人の視覚と対応付けると，畳み込みネットワークが単純特徴抽出を並列に行う前注意過程に相当し，attention による部分特徴の読み出しと全結合ネットワークが，単純特徴を組み合わせて処理する集中的注意過程に相当する．Attention network は特徴マップを入力に持ち，特徴マップと同じ空間サイズの attention をボトムアップに生成する．人の視覚的注意はボトムアップな要因に加え，目的や記憶に依存するトップダウンな要因も持つが，本提案ではトップダウンな視覚的注意のモデル化は行わない．部分的な特徴の読み出しは，生成した attention と特徴マップ間で行列積を取ることで実現される．読み出した部分特徴は全結合ネットワークに入力され，物体のクラスの推定に使われる．

3.1.1　Attention Network

まず，attention network の例を図 3.2 に示す．Attention network の目的は，画像中の物体の位置や大きさを捉え，識別に必要な領域を表す attention を生成することである．畳み込みネットワークから出力された特徴マップを $U \in \mathbb{R}^{H \times W \times C}$ とする．ここで H，W は特徴マップの縦と横の大きさ，C はチャンネル数を表す．Attention network はこの特徴マップ U を入力とし，attention$A \in \mathbb{R}^{H \times W \times V}$ を生成する．V は生成する attention の本数を表す．

Attention network は画像全体の情報を集約した上で注目領域を決める為に，必ず全結合層を含む．しかし，空間解像度を維持したまま全結合を行うとパラメータ数を大きくなる

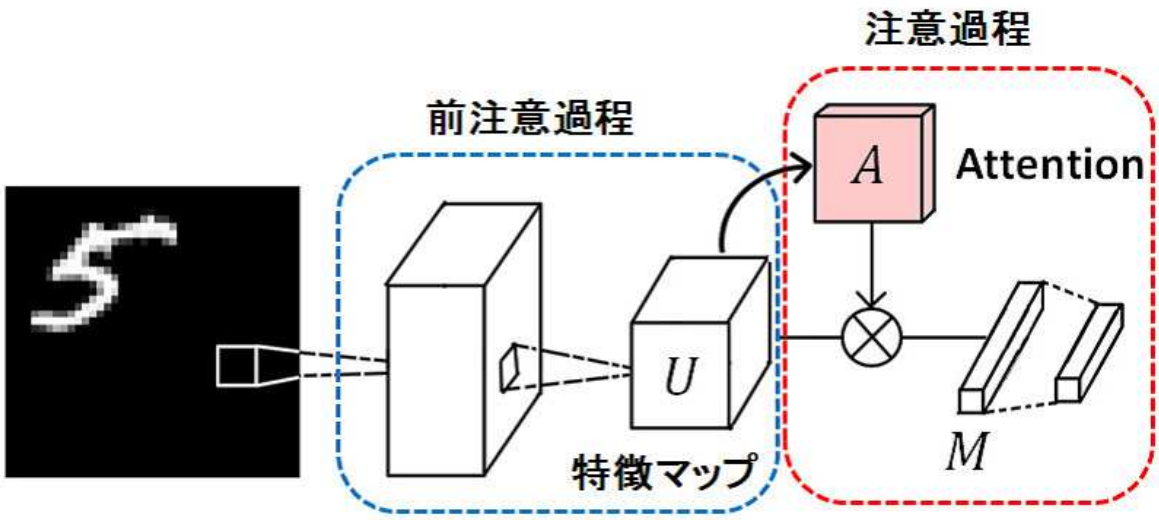

図 3.1: ACNN 概要

ため，畳み込みよって空間解像度を落としてから全結合を行う．特に全結合層のボトルネック部分には，入力画像の空間的な情報が集約されている事が期待でき，この部分を空間特徴量と呼称する．

Attention network は物体の配置の特定のみを目的とし，種類の特定までは行わない為，畳み込み層が持つフィルタの次元は特徴マップよりも小さくて良い．図 3.2 では，64 次元ある特徴マップを，最初の畳み込みによって 3 次元まで圧縮している．この次元圧縮は，attention network のパラメータ数を減らし，計算効率を高める上で重要である．

最後に逆畳み込みによって，空間特徴量から特徴マップ U と同じ大きさ H，W をもつ attention A を生成する．この時，A は sigmoid 関数によって [0〜1] の範囲に正規化を行う．

3.1.2 Attention による特徴抽出と識別

図 3.3 のように空間の次元 H と W を 1 次元に展開した特徴マップ $U \in \mathbb{R}^{\mathrm{HW} \times C}$，及び attention $A \in \mathbb{R}^{\mathrm{HW} \times V}$ を用い，次式によって特徴 $M \in \mathbb{R}^{\mathrm{C} \times V}$ を抽出する．

$$M = U^T \otimes A \tag{3.1}$$

ここで，$\otimes$ は行列積を表す．この行列積により，atttention を空間方向の重みとして特徴マップから特徴量が抜き出される．M の大きさは，U と A のそれぞれチャンネルの次元の積 $\mathrm{C} \times V$ となり，元となる特徴マップの空間の大きさには依存しない．また，離散的

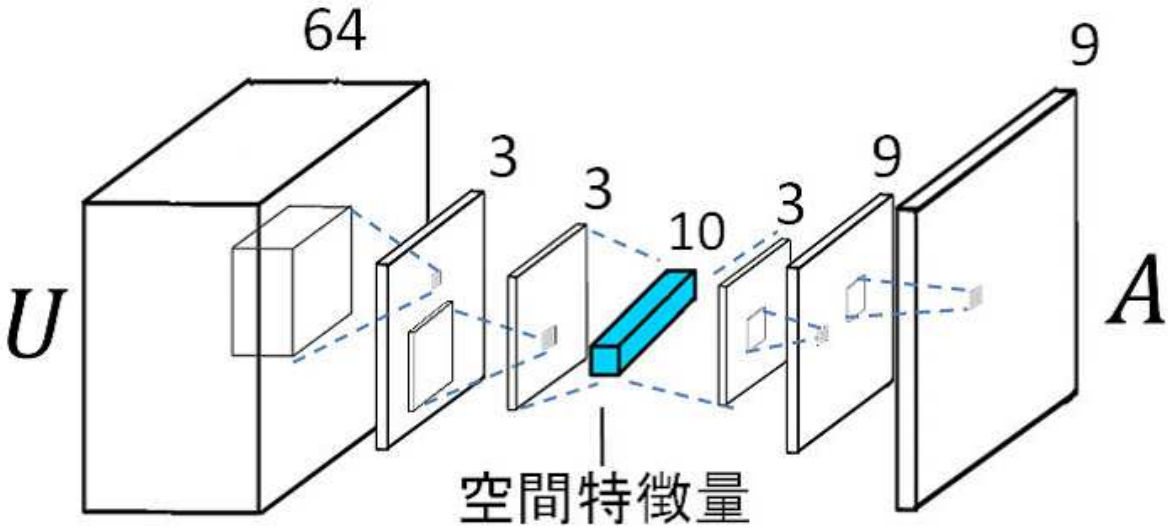

図 3.2: Attention network
各層の上の数字はチャンネル数を表す．カーネルサイズは Encoder 側の畳み込み層は (9 × 9)，decoder 側の逆畳み込み層は (3 × 3)．Stride は全て (2 × 2)．中心のボトルネック部分を空間特徴ベクトルと呼ぶ．

な画素のサンプリングによって局所的な特徴を抜き出す既存手法 [5,6] と異なり，単純な行列積を用いて特徴量を抜き出すので，線形補完等の操作を加えることなく微分可能である．また，attention をニューラルネットの重みとして解釈すると，入力データに応じてモデルの構造が変化するメタネットワーク [22] としての側面も持つ．

識別を行う為に，特徴 M を 1 次元に展開し，全結合層からなる識別用のネットワークに入力する．全結合層の最終層は識別するクラス数と同じユニットを持ち，活性化関数として softmax を用いる．学習に用いる損失関数 $loss$ は，この最終層と正解ラベルとの予測誤差に加え，A に対する正則化項として，A の L1 ノルムを用いる．この時，A の L1 ノルムには $1e^{-5}$ 程度の小さい係数を与える．

3.1.3 再構成誤差

更に，Hinton らの研究 [7] を参考に，付加的な制約項として再構成誤差を用いる．図 3.5 のように，attention network の空間特徴量と，ラベルの予測結果を元に入力画像の再構成を行い，入力画像との再構成誤差を $loss$ に加える．この制約項によって，再構成に必要な物体の配置に関する完全な情報が空間特徴量に保存されることを期待する．

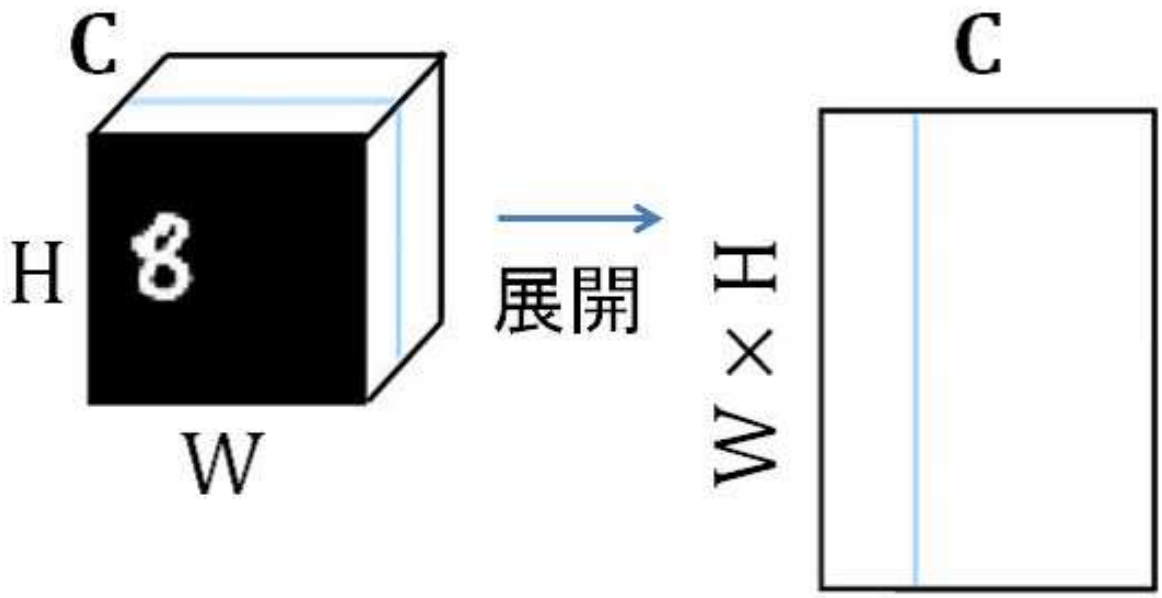

図 3.3: 特徴マップの行列化

再構成ネットワークの構成例を図 3.4 に示す．入力にはラベルの予測結果と空間特徴量の要素積を取ったものを用い，逆畳み込みによって入力画像と同じ大きさのマップを生成する．最終層の活性化関数に sigmoid を用いて出力を正規化し，再構成画像とする．再構成誤差の値は，入力画像と再構成画像の画素毎にクロスエントロピーを取り，画素全体で平均を取ったもの用いる．再構成ネットワークの学習は，他のネットワークと同時に行う．

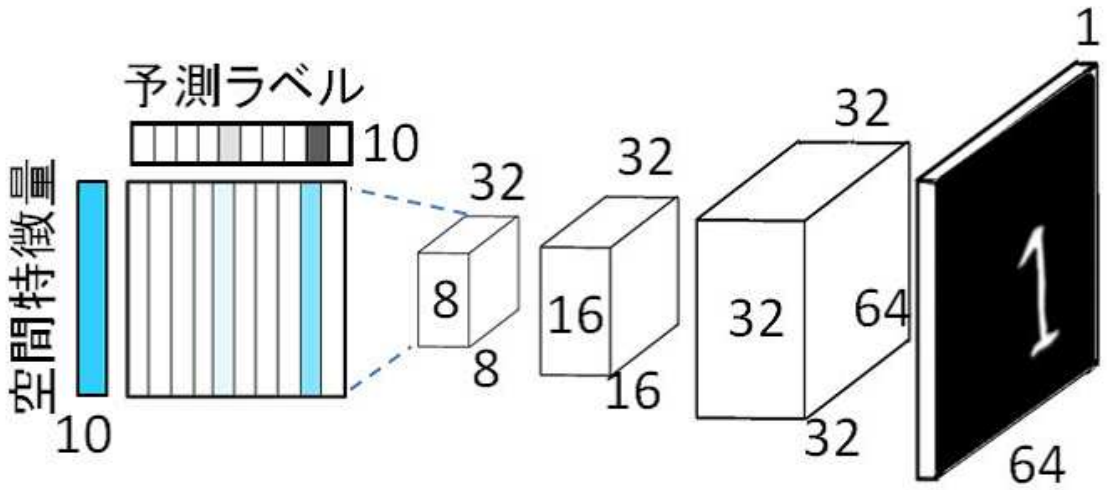

図 3.4: 再構成ネットワーク

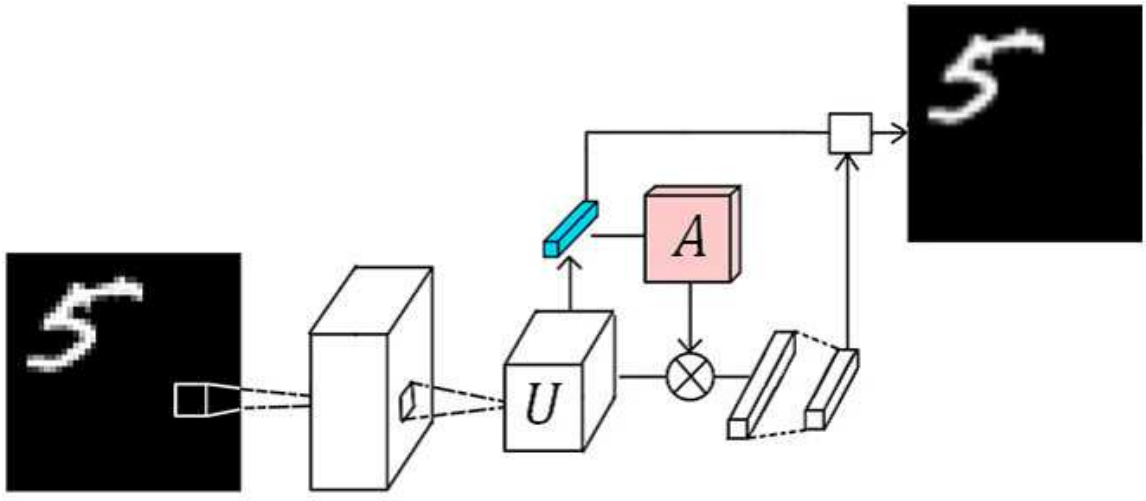

図 3.5: 再構成ネットワークを加えた ACNN
青く塗った層は空間特徴量を表す.

第4章　評価実験

実験1では，平行移動を加えた手書き数字データセットによって学習を行い，大きな物体の位置変化を含むデータセットに対するACNNの性能を通常のCNNと比較する．また，データが少量の場合の過学習の起こり易さと，学習用データに加えていない未知のアフィン変換が与えられた場合の汎化性能も合わせて検証する．

実験2では，空間特徴量が物体の位置に関する情報を保存していることを確認する．なお，実験に使用した各モデルの実装には，全てtensorflow[1]を用いた．

4.1　実験1 平行移動を加えたデータセットに対する性能の検証

4.1.1　実験設定

本実験では，平行移動を加えた手書き文字データセットのMNIST [8]を用い，数字の種類をモデルに識別させる．実験に用いる学習用データの条件毎に，実験を3つに分けた．

実験1.1ではMNISTの5万枚の全ての学習用データを用い，加える平行移動はstep毎に新たにランダムに生成したものを用いた．これは，一般的にデータ拡張として用いられる方法である．

実験1.2では，5000枚のみを学習用データとしてMNISTから切り出し，更に平行移動は学習前にデータに加えた．実験1.1と異なり，学習全体を通して平行移動は各数字に対して1 パターンしか与えられない．

実験1.3では，実験1.1と同様の設定で学習を行うが，性能のテストには公開データセット affNIST[2]を用いて行う．affNISTは平行移動に加えて，回転，拡大縮小，せん断をランダムに加えたMNISTに加えたデータセットである．

平行移動は，黒背景のランダム位置に数字を置くことによって生成した．黒背景の大きさは，実験1.1，および実験1.2では64×64のサイズとした．オリジナルのMNISTの数字の大きさは28×28であるので，上下左右に最大± 18の範囲でランダムに平行移動が加

[1] https://www.tensorflow.org/
[2] http://www.cs.toronto.edu/~tijmen/affNIST/

えられる．実験 1.3 では affNIST データの大きさである，40× 40 のサイズで学習を行った．

比較を行う CNN は，畳み込み層と max-pooling 層を 2 回繰り返した後，2 つの全結合層に繋いだものを用意した．畳み込み層のカーネルサイズは 3× 3，チャンネル数はそれぞれ 32 と 64，max-pooling 層の stride は 2 とした．また，全結合層のユニット数はそれぞれ 128, 10 とした．

次に ACNN は，2 回の畳み込み層を経たものを特長マップ U とし，図 3.2 の attention network によって特徴 M を抽出後、2 層の全結合層に接続した．畳み込み層のカーネルサイズは 3× 3，チャンネル数はそれぞれ 32 と 64，stride はそれぞれ 1 と 2 とし，max-pooling 層は含まない．全結合層のユニット数はベースラインと同様に，それぞれ 128, 10 とした．生成する attention の数 V は 9 枚とした．また，再構成誤差を正則化項として用いるモデル (ACNN-R) は再構成ネットワークを追加して学習を行った．

活性化関数は softmax を用いる全結合層の最終層，及び sigmoid を用いる attention network と再構成ネットワークの最終層を除き，両モデルで共通して relu を用いた．全ての実験を通して optimizer には adam を使用し，学習率は 0.001 で固定した．また，batch の大きさは 100 とした.

4.1.2　実験 1.1 の結果

100epoch の学習を行った後のテスト用データに対する accuracy，および学習に使用したパラメータ数を表 4.1 に示す．また，学習後に生成された ACNN と ACNN-R の attention を図 4.1，図 4.2 に示す．また，attention のチャンネル毎に異なる彩色を施したものを図 4.3 に示す.

表 4.1: 平行移動を加えた MNIST に対する accuracy

method	test accuracy (%)	パラメータ数
CNN (baseline)	99.02	2117k
ACNN	99.21	115k
ACNN-R	**99.25**	202k

4.1.3　実験 1.2 の結果

学習用データを少量にした条件で，100epoch の学習を行った後のテスト データに対する accuracy を表 4.2 に示す．なお，train データに対する accuracy は全てのモデルで 100%となった．各モデルの学習中の train と test の accuracy の変化を，図 4.4，4.5，4.6 に示す．

4.1.4　実験 1.3 の結果

モデルにとって未知のアフィン変換を含む affNIST の test データ 32 万枚に対する正解率を表 4.3 に示す．なお，平行移動のみを加えたテスト データに対する正解率は，全てのモデルが 99%を超えている．

4.1.5　考察

表 4.1 より，ACNNR と ACNN-R は共に CNN よりも高い識別性能を示している．後段の全結合層の構成は全てのモデルで同一である為，attention network を用いた特徴抽出はより平行移動により頑健であると言える．今回の実験設定では数字の移動範囲が ±18 あるのに対し，ベースラインとなる CNN の全結合層の直前の pooling 層が持つ受容野の広さは 10× 10しかなく，ベースラインの CNN では位置の変動を吸収しきれなかったと考えられる．

次に，図 4.1 と図 4.2 を見ると，生成された attention は教師データとして数字の位置を与えていないにも関わらず，数字の位置に集まっている．識別に必要な情報を全結合層に集める為に，attention network が数字の位置の捉え方を学習したと考えられる．また，ACNN と ACNN-R の attention を比較すると，再構成誤差を加えた ACNN-R の方が数字全体を覆う attention が生成されやすい傾向が見られた．再構成を行う為には，数字の大きさや傾き

表 4.2: 少量データに対する accuracy

method	test accuracy (%)
CNN (baseline)	79.45
ACNN	92.07
ACNN-R	**94.28**

といった属性の情報が必要であるため，数字全体を見るよう学習が進んだと考えられる．

更に，図 4.3 のチャンネル毎に彩色を施した attention を見ると，ACNN-R から生成された attention には，入力画像に関係なく層状の構造が安定して見られる．Attention の構造が安定しているという事は，後段の全結合層への入力が安定していることを意味し，ACNN-R の正解率が ACNN を上回る結果に繋がったと考えらえる．

次に，各モデルのパラメータ数を比較すると，ACNN はベースラインの CNN の 6%，ACNN-R は 10%程度と大幅に抑えらえている．これは最もパラメータを必要とする畳み込み層と全結合層の繋ぎ目の部分のネットワークが，ACNN では attention によって小部分のみを繋ぐよう置き換わった為である．ACNN も attention network 内に全結合層を含んでいる為，計算量は CNN と変わらず画像の面積に比例して増加するが，attention network のフィルタの次元数が十分小さい為にこの影響は小さい．

表 4.2 より，CNN と比較して ACNN は 10%以上高い acuuracy を維持した．図 4.4，4.5，4.6 の学習中の acuuracy の推移を見ると，CNN は早い段階から train と test の間で accuracy に大きな差が生じているが分かる．一方，ACNN，及び ACNN-R も 10epoch 付近から過学習が始まっているが，その開きは CNN と比較して緩やかである．

学習用データが少量である場合には，画像内の物体の位置とそのラベルに相関が無くても，位置とラベルの組み合わせを全結合ネットワークが記憶してしまうことで過学習が起こる．一方，ACNN では，attention network が物体の位置，全結合層が物体のラベルというように別々のネットワークで学習する為，過学習を防ぐことができたと考えられる．

表 4.3 より，未知のアフィン変換に対してもベースラインの CNN と比較して ACNN と ACNN-R は共に高い正解率を維持した．オリジナルの手書き数字が自然に含む小さな回転や大きさのバラエティから，これらに影響を受けにくい attention の配置を attntion network が学習したものと考えられる．

以上より，物体の大きな位置変化に対し，ACNN は attention を動かす事で頑健に識別

表 4.3: 未知のアフィン変換に対する accuracy

Method	affNIST accuracy (%)
CNN	77.62
ACNN	85.33
ACNN-R	**87.55**

を行う事ができると言える．

4.2　実験 2: 同一の空間特徴量による異なる数字の再構成

本実験では，同じ画像から抽出した空間特徴量を異なる数字のラベルと組み合わせても，数字の再構成できるかを検証した．再構成を行うネットワークは，実験 1.1 の実験設定で学習済みの ACNN-R を用い，予測ラベルは各数字を one-hot に変換したベクトルを与えた．再構成結果を図 4.7 に示す．

図 4.7 を見ると，数字の種類が与えたラベルと同一で，位置は空間特徴量の元となった画像と同じ位置に数字が再構成されていることが分かる．この事から，空間特徴量には画像中の物体の位置情報が保存されていることが確認できる．

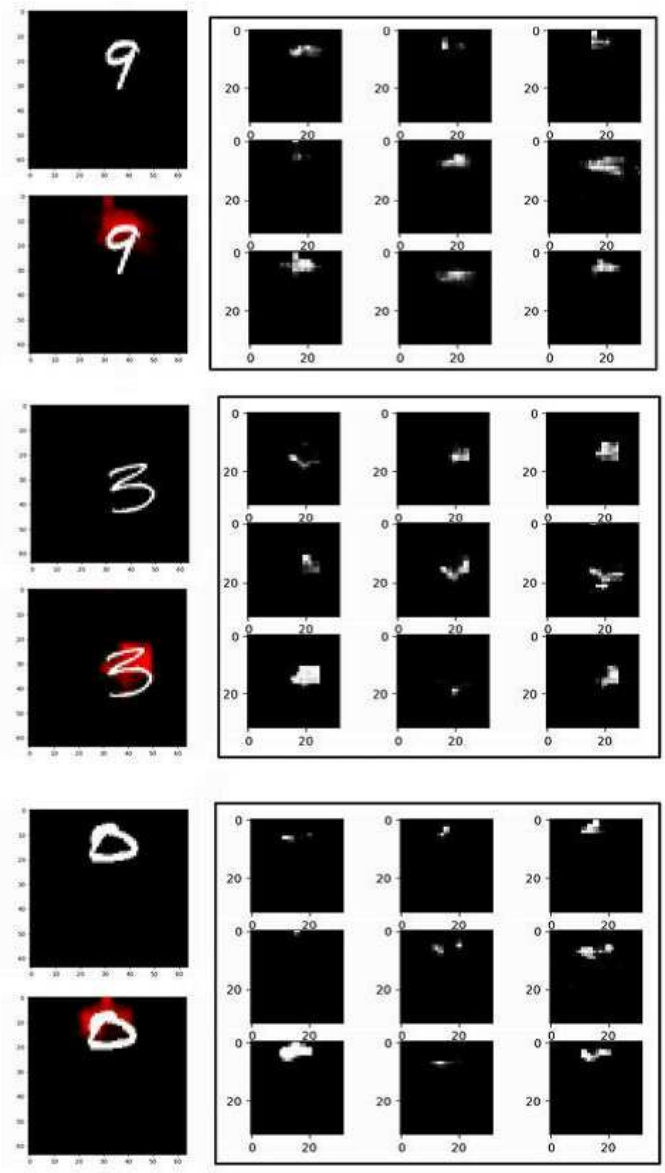

図 4.1: ACNN の attention
左上は入力画像，右の 9 枚の画像は attention の各チャンネルの重みを [0〜1] に正規化した結果を表す．左下の画像は全ての attention を足し合わせた上で赤く彩色し，入力画像に重ねた結果を表す．

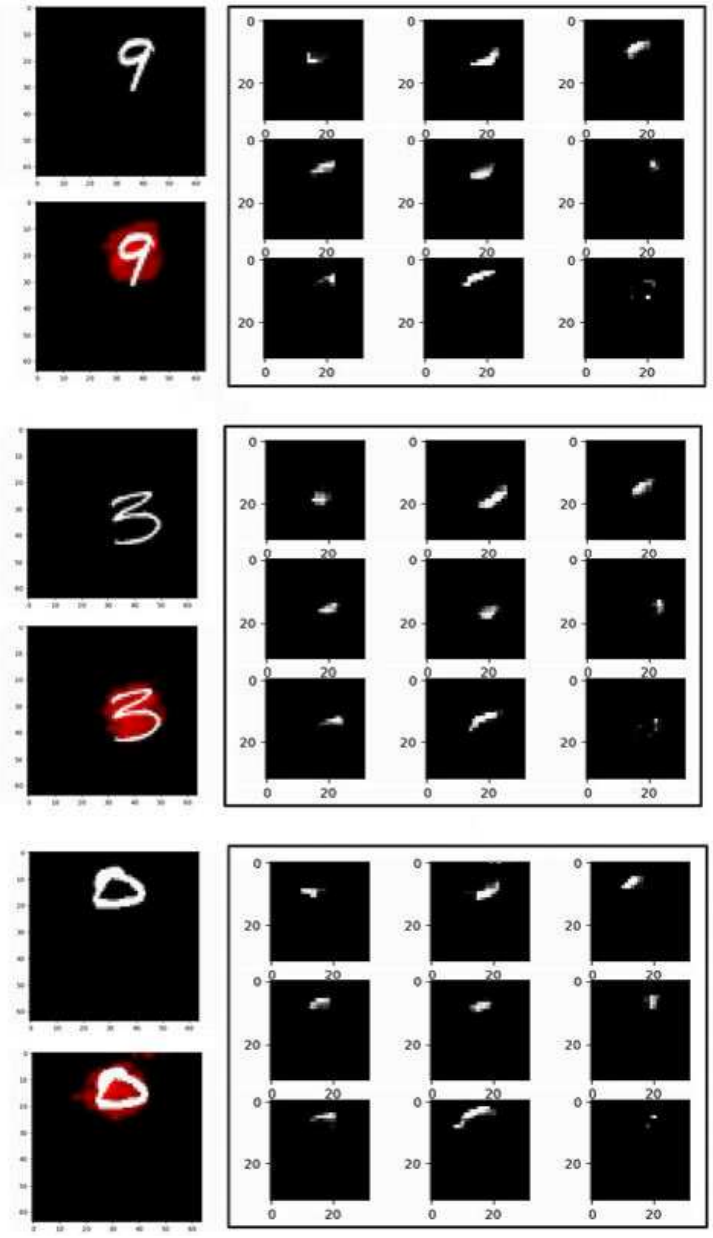

図 4.2: ACNN-R の attention

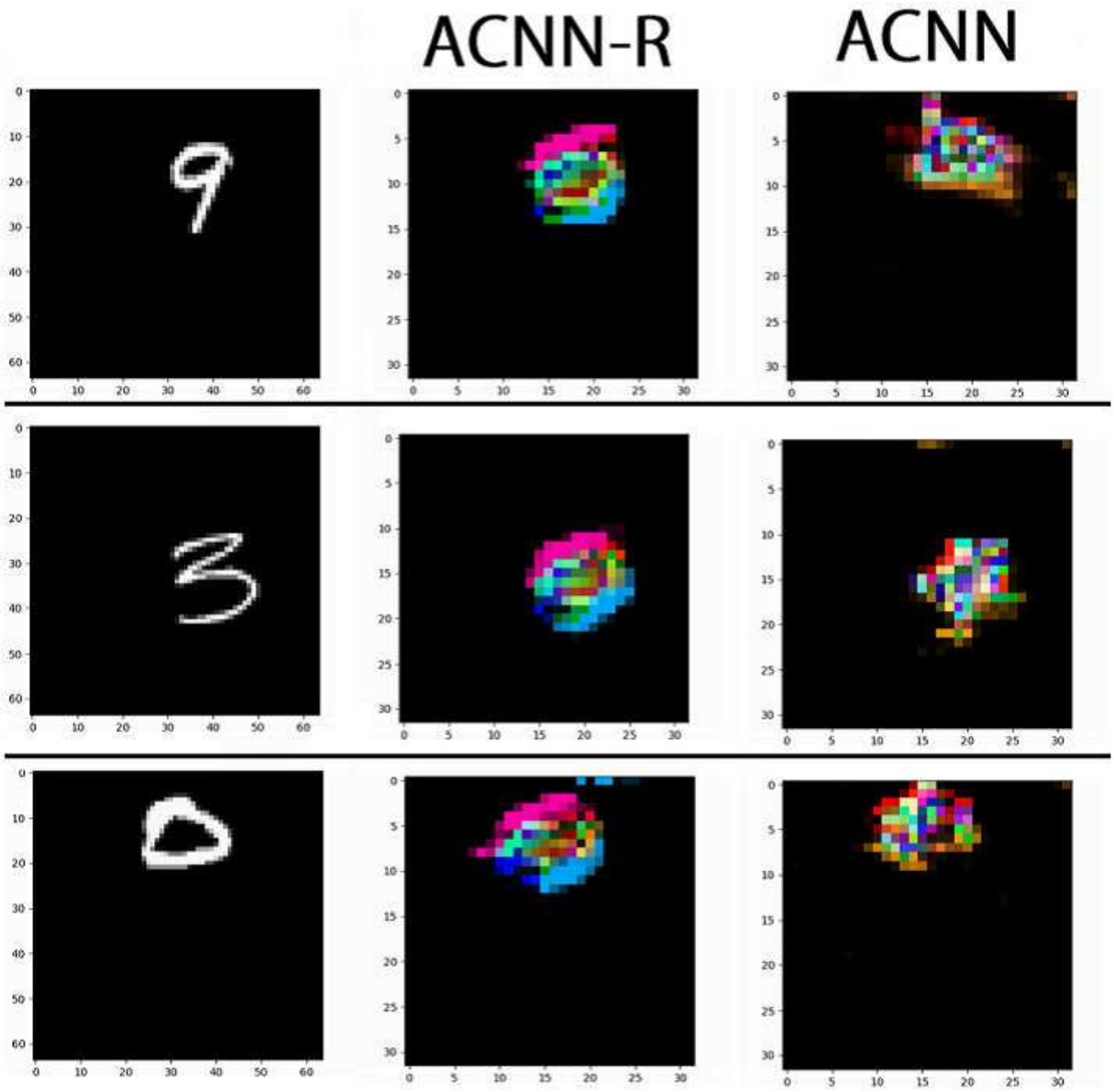

図 4.3: チャンネル毎に彩色した attention
図 4.1 と図 4.2 の 9 枚の attention にそれぞれ異なる色を付けた上で一枚に重ねて作成した.

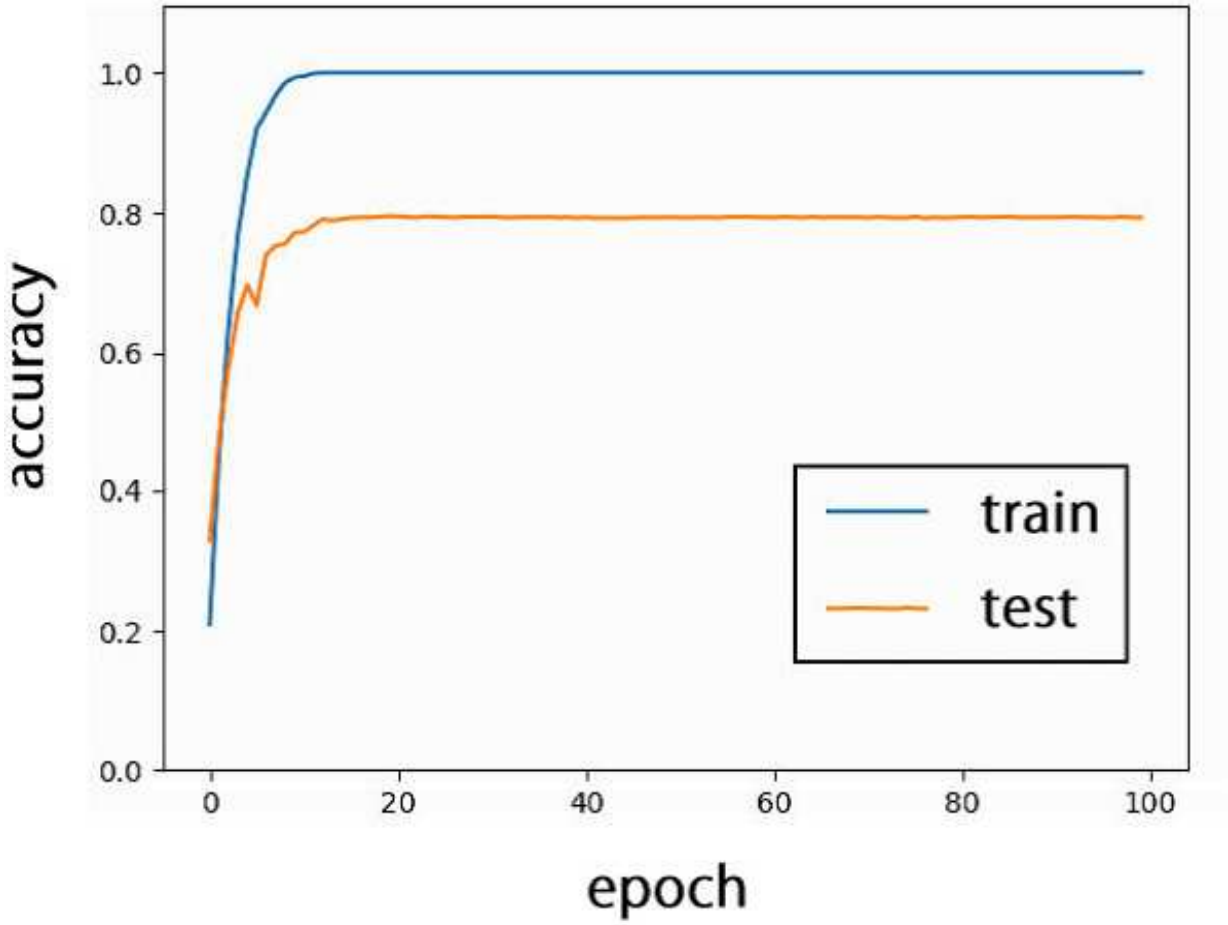

図 4.4: ベースライン CNN の train と test accuracy の推移

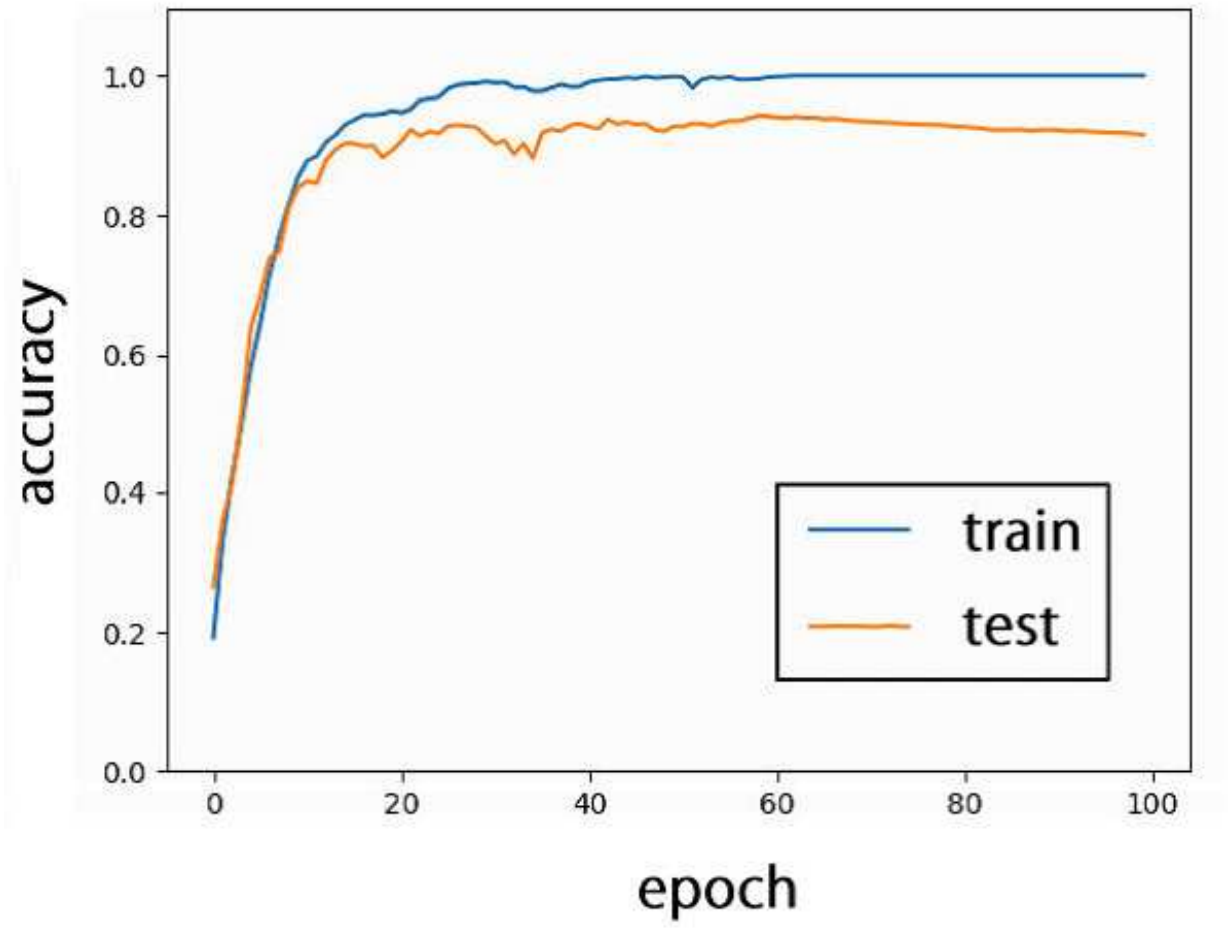

図 4.5: ACNN の train と test accuracy の推移

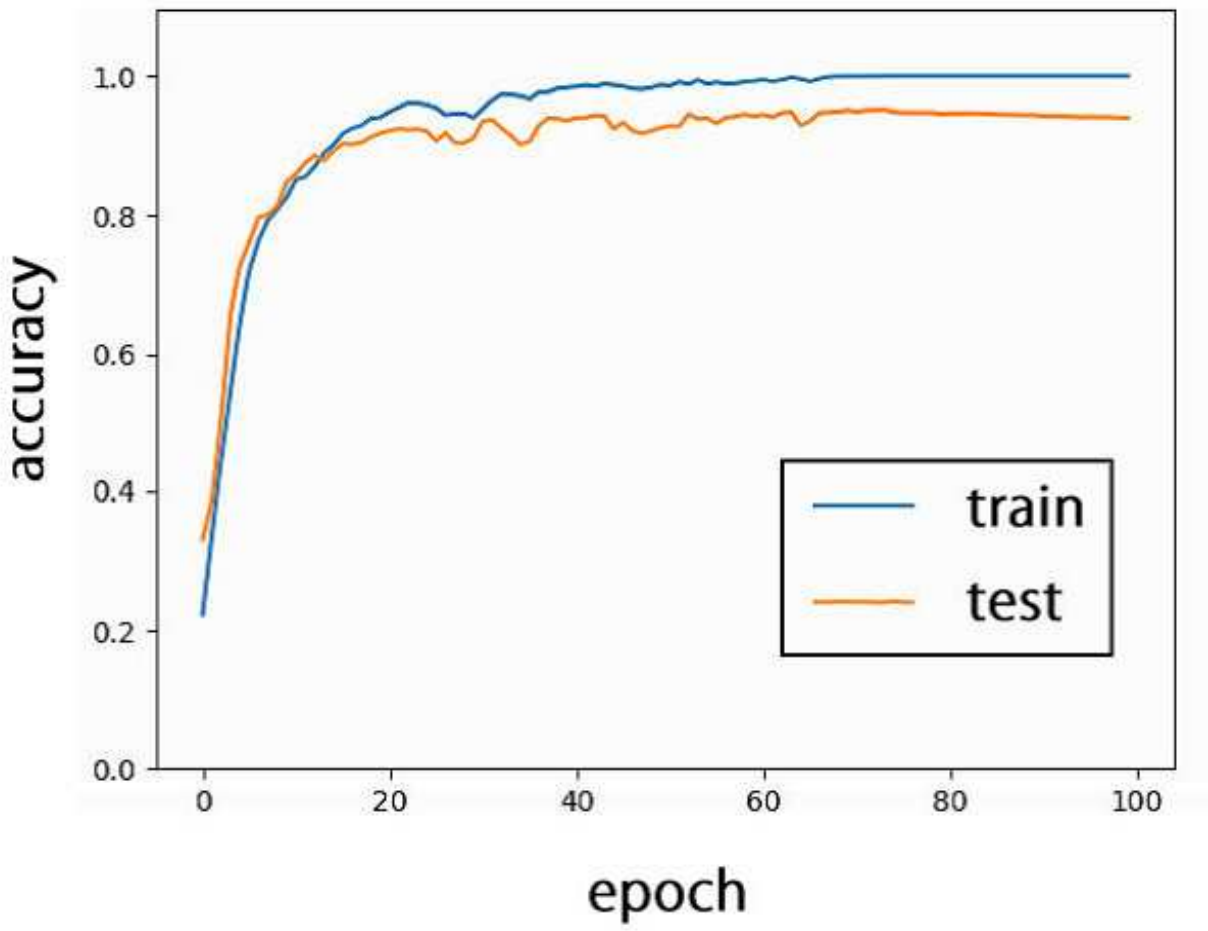

図 4.6: ACNN-R の train と test accuracy の推移

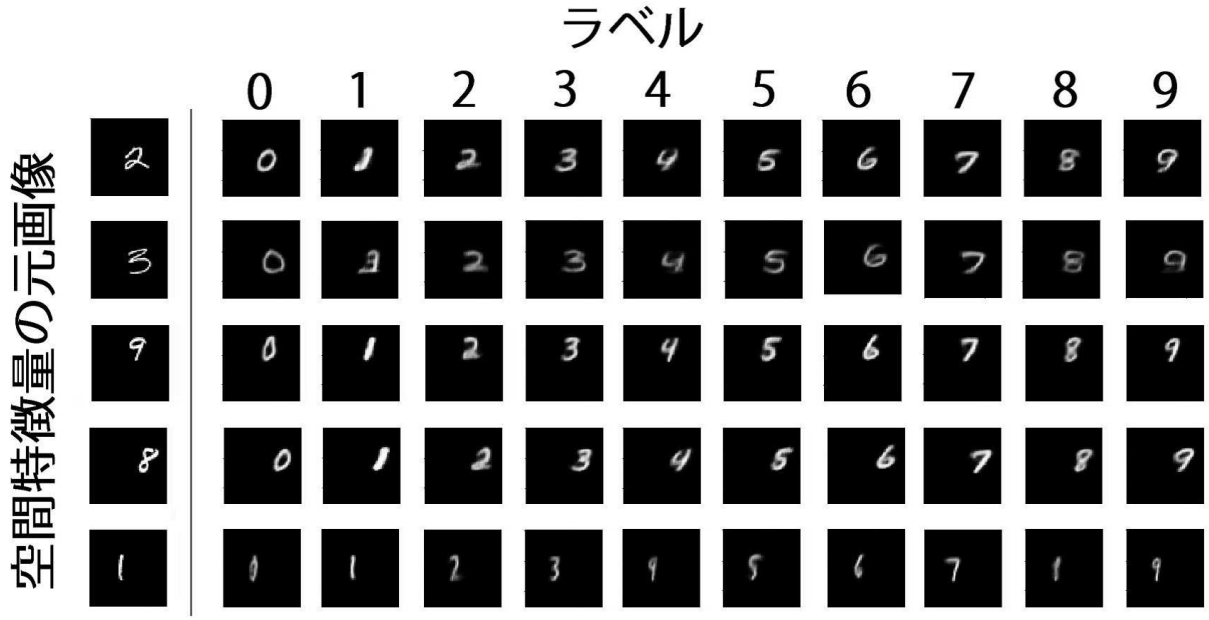

図 4.7: 同一の空間特徴量による異なるラベルの再構成
各行が再構成に使用した空間特徴量，各列が使用したラベルを表す.

第5章　結論

5.1　まとめ

本論文では人の視覚的情報処理の過程を参考に，CNN に attention の構造を取り入れた Attentive Convolutional Neural Network(ACNN) モデルを提案した．従来の CNN の pooling 層に替わって attention を生成する attention network を用いる事で，大域的な位置ずれに頑健になることを確認した．また，空間特徴量に対する付加的な制約項として再構成ネットワークによる再構成誤差が用いると，より汎化性能が上げられる事を確認した．

実験では，複数の条件でアフィン変換を加えた手書き文字データセット MNIST の識別性能を CNN と比較した．実験1 では，平行移動をデータ拡張として MNIST に加えたデータセットに対する性能を CNN と比較し，1 割未満のパラメータ数で同等以上の識別率を得た．計算資源の観点から，ACNN は物体の平行移動をより効率的に学習可能なモデルであると考えられる．また，学習データが少量である条件でも学習を行い，ACNN が過学習を起こし難いモデルである事を確認した．更に，学習時に与えられていない未知のアフィン変換に対する性能も検証し，高い汎化性能を持つ事を確認した．最後に実験 2 では，画像中の物体の位置に関する情報が，attention network に保存されている事を確認した．

5.2　課題と展望

本研究は黒背景に合成された手書き数字という特殊な条件下での実験しか行っておらず，一般物体認識のような複雑な背景を持つ条件でも，同様に attention が有効に働くかについては更なる検証を必要とする．背景が複雑になると，attention network の構造もそれに応じて複雑にする必要があると考えられる．

実世界において起こりうるアフィン変換以外の膨張収縮や部分的な変形といった複雑な設定下でのモデルの頑健性も未検証である．Jaderberg ら研究 [6] と異なり，ACNN は受容野の形状や密度を柔軟に変更できる．この性質は，複雑な形状変化に対して有効に働く事が期待できる．

また，画像中に複数の識別対象があるような状況を考えた時，attention を向ける先を動

的に制御する仕組みが必要である．これに対しては，外部メモリを扱うネットワーク [23,24] を参考に，attention を制御する RNN 式のコントローラの導入を検討している．

謝辞

本研究の着手及び方針について，多くの御指導，御助言を頂いた小林 哲則教授に，心より感謝申し上げます．

また研究に関し，多くの御提案，御助言を頂いた，藤江 真也氏，小川 哲司氏，俵 直弘氏に深く感謝致します．

最後に，研究生活の中で，多くの議論に付き合ってくれた菊池 康太郎氏，赤川 優斗氏，金田 健太郎氏，研究室の皆さまに深く感謝致します

参考文献

[1] A. Treisman, “A feature-integration theory of attention,” Cognitive Psychology, vol.12, no.1, pp.97–136, 1979.

[2] D. Hubel, T. Wiesel, “Receptive fields of single neurones in the cat ’ s striate cortex,” The Journal of physiology, vol.160, pp.106–154, 1962.

[3] B. Cheung, E. Weiss, and B. Olshausen，“The Emergence of a Fovea while Learning to Attend，” International Conference on Learning Representations，2017.

[4] Y. Jeon, J. Kim，“Active Convolution: Learning the Shape of Convolution for Image Classification，” Computer Vision and Pattern Recognition，pp.1846–1854，2017.

[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y Wei, “Deformable Convolutional Networks ,” International Conference on Computer Vision, pp.764–773, 2017.

[6] M. Jaderberg，K. Simonyan，A. Zisserman，K. Kavukcuoglu，“Spatial Transformer Networks，” Advances in Neural Information Processing Systems，pp.2017–2025，2015.

[7] S. Sabour, N. Frosst, and G. E Hinton，“Dynamic Routing Between Capsules，” Advances in Neural Information Processing Systems，pp.3859–3869，2017.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner，“Gradient-based learning applied to document recognition，” Proceedings of the IEEE，pp.2278-2324, 1998.

[9] Z. Cao, T. Simon, S. Wei, Y. Sheikh，“Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields，” Computer Vision and Pattern Recognition, 2017.

[10] K. Lenc and A. Vedaldi，“Understanding image representations by measuring their equivariance and equivalence，” Computer Vision and Pattern Recognition, pp.991–999, 2015.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and et.al, “Going Deeper with Convolutions, ” Computer Vision and Pattern Recognition, pp.1–9, 2015.

[12]] E. Shelhamer, J. Long and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation, ” Pattern Analysis and Machine Intelligence, vol.39, no.4, pp.640–651, 2017.

[13] A. Radford, L. Metz, S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, ” arXiv, 2015.

[14] J. Zhu, T. Park, P. Isola, and A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, ” International Conference on Computer Vision, 2017.

[15] M. Cheng, Z. Zhang, W. Lin, and P. Torr, “BING: Binarized Normed Gradients for Objectness Estimation at 300fps, ” Computer Vision and Pattern Recognition, pp.3286–3293, 2014.

[16] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders, “Selective Search for Object Recognition, ” International Journal of Computer Vision, vol.104, no.2, pp.154–171, 2013.

[17] R. Girshick, J. Donahue, T. Darrell, J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation, ” Computer Vision and Pattern Recognition, pp.580–587, 2014.

[18] R. Girshick, “Fast R-CNN, ” International Conference on Computer Vision, pp.1440–1448, 2015.

[19] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, ” Advances in Neural Information Processing Systems, pp.91–99, 2015.

[20] J. Redmon, S. Divvala, R. Girshick and A. Farhadi，“You Only Look Once: Unified, Real-Time Object Detection，” Computer Vision and Pattern Recognition, pp.779–788, 2016.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, Cheng-Yang Fu, A. C. Berg，“SSD: Single Shot MultiBox Detector，” European Conference on Computer Vision, pp.21–37, 2016.

[22] D. Ha, A. Dai, Q. V. Le，“HyperNetworks，” International Conference on Learning Representations, 2016.

[23] G. Edward，and et.al, “Hybrid computing using a neural network with dynamic external memory，” Nature(journal), vol.538, pp.471-476, 2016.

[24] S. Sainbayar, S. Arthur, W. Jason, F. Rob, “End-To-End Memory Networks，” Advances in Neural Information Processing Systems, pp.2440–2448, 2015.