Visual Question Answering のための画像認識手法の考察

1　はじめに

近年，機械学習が注目されており，様々な分野において研究が為されている．最近では，画像や言語，音声を複合させたマルチモーダルな機械学習の研究も報告されており，人工知能が画像処理や自然言語処理といった個々の分野を超えたより複雑な問題への応用が期待されている．マルチモーダルな研究の一例として，Visual Question Answering (VQA) が挙げられる．これは画像と，その内容について尋ねる自然言語文での質問が与えられたとき，適切な回答を予測するタスクである．VQA は 画像処理と自然言語処理を組み合わせた研究分野であり，今後の発展が望まれている．. 今回の実験では，マルチモーダルデータを扱う前段階として，丸，三角，四角を含む画像を対象とした実験をする．モデルや各種パラメータの差が識別結果にどのような影響を与えるかを調査し，画像処理への理解を深める．

2　Visual Question Answering (VQA)

VQA は，ある画像とその画像に関する質問を入力として受け取り，自然言語による正しい回答を導き出すタスクのことである．図 1 に VQA に対する機械学習の適用の一例として，Agrawal が提案したモデルを示す [1]．このモデルでは画像を処理する CNN の出力とテキストを処理する LSTM の出力を掛け合わせてマルチモーダル表現空間を形成している．

3　Natural Language for Visual Reasonin (NLVR)

VQA の一種である NLVR は，色の付いた図形が複数描かれた画像とキャプションのセットが与えられたとき，そのキャプションが正しく図を説明しているかを True/False で判定するタスクである．図 2 に NLVR のデータの一例を示す．このデータには，”There is exactly one black triangle not touching any edge.” という図を正しく説明したキャプションが与えられているため，答えは True となる．

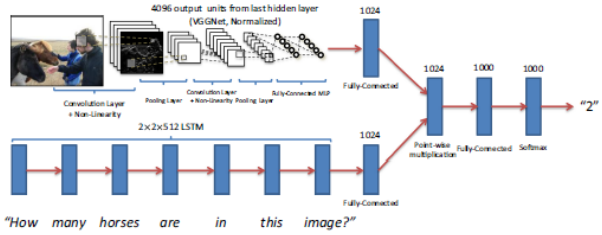

図 1: VQA のためのモデルの一例 [1]

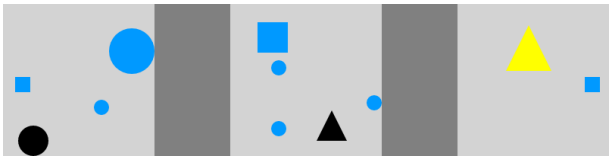

図 2: NLVR データセットの例

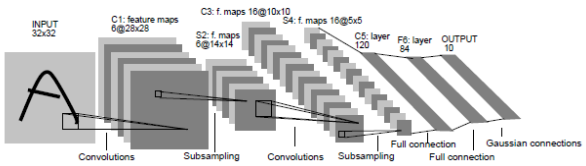

図 3: Convolutional Neural Network モデルの一例 [2]

4　要素技術

4.1　Convolutional Neural Network (CNN)

CNN は，単純な順伝播型のニューラルネットワークとは違い，全結合層だけでなく畳み込み層 (Convolutional Layer) とプーリング層 (Pooling Layer) から構成されるニューラルネットワークである．CNN では畳み込み層とプーリング層を通すことでニューラルネットワークでの学習が効率的になるように入力を変換する．畳み込み層では，入力データに対してカーネルと呼ばれる小さな行列をスライドさせながら適用していく．入力データの一部に対して対してカーネル積和を計算し，カーネルの位置をずらしながら新しい行列を作成していく．この新しい行列を特徴マップと呼ぶ．単純なニューラルネットワークでは入力データのサイズが大きくなるほど重みパラメータ数が増大していく．しかし畳み込み

図 4: 学習画像

層では，入力データのサイズが大きくなったとき，特徴マップのサイズは増えるがカーネルのサイズ，つまり重みパラメータ数は変化しないという特徴がある．プーリング層では画像を縮小することで小さな位置変化に対して頑強になる．最もよく使われる max pooling では，入力データを小さな領域に分割し，各領域の最大値をとってくることでデータを縮小する．データが縮小されるため，計算コストが軽減されることに加えて，各領域内の位置の違いを無視するため，小さな位置変化に対して頑健なモデルを構築することができる．このように画像のフィルタ処理におけるフィルタの値を学習し，特徴を抽出する．

4.2 VGG16

VGG16 とは，畳み込み 13 層と全結合 3 層の計 16 の隠れ層からなる畳み込みニューラルネットワークである．Oxford 大学の VGG (Visual Geometry Group) が提案し，2014 年の ILSVRC (ImageNet Large Scale Visual Recognition Challenge) で好成績を収めた．

5 実験

本実験では，NLVR データセットから図形を切り出した画像を入力とし，その図形の形を識別する．NLVR データセットには circle, triangle, square の 3 種類の図形が含まれており，図形は black, blue, yellow の 3 種類の色の内いずれかである．また，学習モデルとして, 3 層からなる単純な Neural Network (NN) モデル，CNN モデル， VGG16 の 3 つを用いた．以下，それぞれを NN モデル, CNN モデル, VGG16 モデルとする．表 1, 表 2，表 3 にこれら 3 つのモデルの詳細を示す．ただし VGG16 モデルには新たに全結合層と出力層を追加し，重みの更新は既存の層の出力層に近い部分と追加した層のみとした．

5.1 実験 1

実験 1 ではモデル構造の違いによる識別率の差を調べた．図 4 に示した図形をそれぞれ 3 つのモデルに 800 件ずつ学習させ，図形の形で 3 クラス識別をした．この時のエポック数は 50 とした．表 4 に学習データ 800 件とテストデータ 200 件を識別させた結果を示す．次に，表 5, 表 6，表 7 にテストデータの図形，色毎の識別結果を示す. 表 4 より NN と CNN を比べると，NN のほうが精度がやや悪いことが分かる．さらに表 5 より，yellow の square の特徴を学習しきれていないことがわかる．誤識別した yellow の square はいずれも画像サイズが 14 ピクセル以下の小さな画像であり，これらを circle と識別していた．このことから，画像サイズが小さいものを入力するときは，単純な拡大ではなく何らかの工夫が必要であると考えられる．実験 1 の結果より，CNN と VGG16 は単純な NN モデルと比べるとより高い精度で画像を識別できるといえる．

表 1: NN モデルの条件

入力サイズ	224*224*3
ドロップアウト	0.50
中間層のノード数	512
活性化関数	ReLU
Optimizer	SGD

表 2: CNN モデルの条件

入力サイズ	(224, 224, 3)
ドロップアウト	0.50
フィルタサイズ	3
プーリングサイズ	2
プーリングタイプ	MaxPooling
活性化関数	ReLU
Optimizer	SGD

表 3: VGG16 モデルの条件

入力サイズ	(224, 224, 3)
ドロップアウト	0.50
全結合層のノード数	512
活性化関数	ReLU
Optimizer	SGD

表 4: 各モデルのテストデータの識別結果

	NN	CNN	VGG16
loss	0.0114	2.63e-06	1.05e-04
accuracy	0.993	1.0	1.0

表 5: NN モデル: 図形，色毎の識別結果

	circle		triangle		square		total	
	loss(e-04)	acc	loss(e-04)	acc	loss(e-04)	acc	loss(e-04)	acc
black	0.00119	1.0	0.00119	1.0	0.0103	1.0	0.00427	1.0
blue	0.879	1.0	0.00653	1.0	0.0586	1.0	0.314	1.0
yellow	278	1.0	0.00861	1.0	778	0.940	352	0.980
total	71.7	1.0	0.00459	1.0	195	0.985		

表 6: CNN モデル: 図形，色毎の識別結果

	circle		triangle		square		total	
	loss(e-07)	acc	loss(e-07)	acc	loss(e-07)	acc	loss(e-07)	acc
black	7.71	1.0	1.19	1.0	128	1.0	45.8	1.0
blue	34.8	1.0	3.25	1.0	44.0	1.0	27.3	1.0
yellow	34.3	1.0	11.5	1.0	75.3	1.0	40.4	1.0
total	22.2	1.0	5.48	1.0	52.1	1.0		

表 7: VGG16 モデル: 図形，色毎の識別結果

	circle		triangle		square		total	
	loss(e-03)	acc	loss(e-03)	acc	loss(e-03)	acc	loss(e-03)	acc
black	0.101	1.0	0.156	1.0	0.119	1.0	0.125	1.0
blue	0.0591	1.0	0.0255	1.0	0.0426	1.0	0.0424	1.0
yellow	0.151	1.0	0.123	1.0	0.133	1.0	0.136	1.0
total	0.123	1.0	0.176	1.0	0.114	1.0		

5.2 実験 2

実験 2 では CNN モデルのハイパーパラメータの値を変更したときの学習の様子について調べた．図 5 にカーネルサイズを 3, 5, 9, 11 と変更したときの学習の推移を，図 6 にプーリングサイズを 2, 8, 32, 64 と変更したときの学習の推移を示す．ここでのエポック数は 20 とした．図 5 より，カーネルサイズが大きいモデルほど精度が悪いことがわかる．この原因として，カーネルサイズが大きいと画像の特徴を抽出できる範囲が大きくなるが今回のような図形の識別といったタスクでは広範囲の特徴抽出をする必要がないということが考えられる．また図 6 より，プーリングサイズが大きいモデルほど収束が遅く，精度が悪いことがわかる．この原因として，プーリングサイズが大きいとプーリングによる情報量の欠落がより大きくなるために一つ一つの画像の識別が困難になることが考えられる．NN モデルとプーリングサイズを変更した CNN モデルを比較する．NN モデルでは学習パラメータの数は 77,072,387 であり，プーリングサイズ 8 と 32 の学習パラメータの数はそれぞれ 11,946,883 と 592,771 であった．表 8 よりプーリングサイズ 8, 32 のモデルのテストデータの正解率は共に 1.0 であることと，表 4 の NN モデルの正解率から，NN モデルに比べて CNN モデルはより少ないパラメータ数で高い学習精度を出せるということがわかる．

表 8: プーリングサイズ 2, 8, 32, 64 におけるテストデータの識別結果

	size 2	size 8	size 32	size 64
loss	1.13e-04	7.41e-04	4.79e-03	0.118
accuracy	1.0	1.0	1.0	0.968

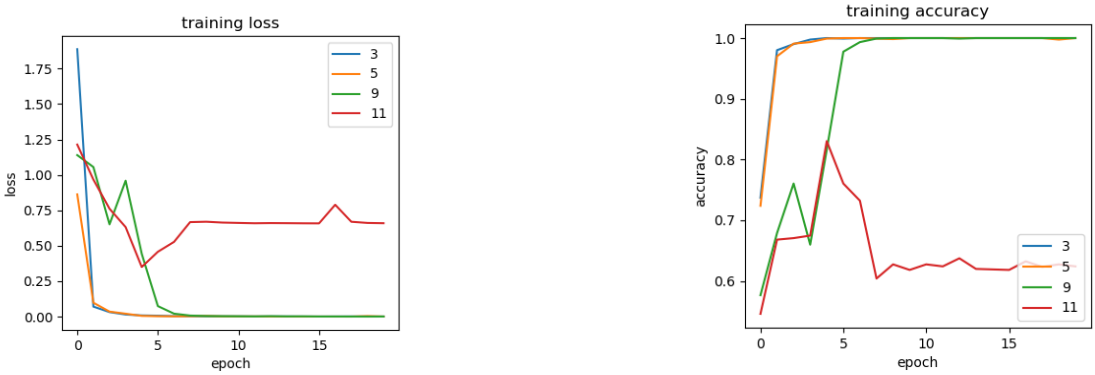

図 5: カーネルサイズ 3, 5, 9, 11 における loss と accuracy の推移

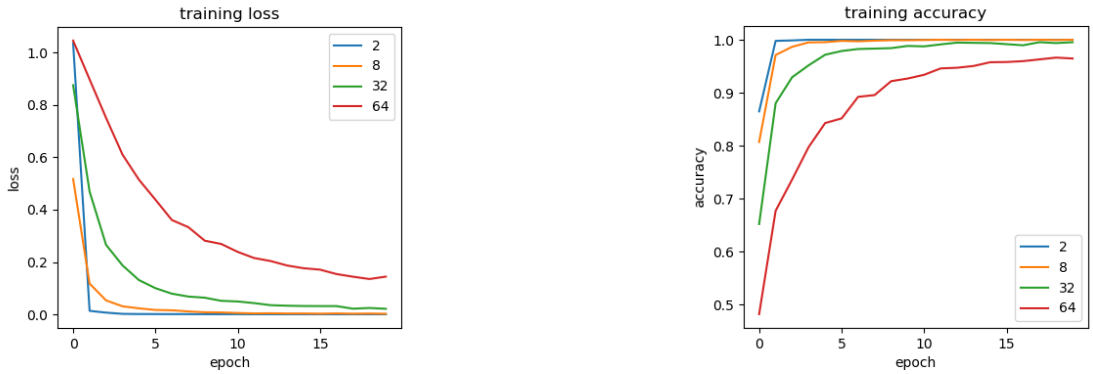

図 6: プーリングサイズ 2, 8, 32, 64 における loss と accuracy の推移

6　まとめと今後の課題

今回，画像とキャプションと質問文から自然言語による回答を得るための準備として 3 種類の図形の識別に関する実験をした．実験 1 の結果から，単純な NN に比べて CNN，VGG16 は画像をより高い精度で識別できると分かった．また入力画像に関する考察から，サイズの小さな画像を正しく識別するためには何らかの工夫が必要であると考えられる．次に実験 2 の結果から，カーネルサイズが大きいと学習精度が低くなるということが分かった．この原因として，今回の図形識別のようなタスクでは広範囲の特徴抽出の必要がないことが考えられる．またプーリングサイズに関する考察から，プーリングサイズが大きいとプーリングによる情報の欠落が大きくなるために学習精度が落ちるということが考えられる．今後の課題としては以下のことが挙げられる．

- 複数の図形が含まれる画像の識別
- 自然言語による応答

複数の図形が含まれる画像を学習させた際の図形識別に関しては，YOLO や SSD などの物体検出アルゴリズムを用いて画像中の図形を取得することを検討する．自然言語による応答に関しては，LSTM を用いて自然言語の処理をし，attention の取り扱いを組み込んだモデルを作成することで，質問文の応答に応じて画像の注目場所を判定することを目標にしたい．

参考文献

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, Vol. abs/1505.00468, , 2015.

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324, 1998.