言語情報を統合して画像中のオブジェクト間の関係推定を行うニューラルネットワーク

黒澤 郁音　　小林 哲則　　林 良彦
早稲田大学理工学術院
ikuto@pcl.cs.waseda.ac.jp

1 はじめに

与えられた画像に対して説明文を生成する研究 [7] が盛んに行われているが，これらは，画像一枚に対して一文の説明文を生成するものがほとんどである．一般に，画像には複数のオブジェクトが含まれており，一文の説明文によって画像に写された情報を記述するのは難しい．これに対し，複数存在する画像中のオブジェクト間の関係を抽出できれば，画像をより詳細に表すことが可能となり，画像に対する意味的なメタデータとして，検索などに有効に利用することができる．

本研究では，画像中のオブジェクトの位置とそのクラスラベルが与えられていることを前提とし，これらの間を結ぶ関係 (predicate) を推定する問題を扱う．この問題に対し，クラスラベルから得られる言語特徴，bounding box から得られる領域特徴，畳み込みニューラルネットワーク (CNN) によって得られる画像特徴を統合して predicate の推定を行うニューラルネットワークを提案する．また，クラスラベルに対して得られる言語情報の有効性を議論する．

2 問題とアプローチ

図 1 にオブジェクト間の関係抽出の例を示す．まず，この例では 2 つのオブジェクト (*person* と *horse*) が識別されている．すなわち，bounding box (赤線で囲んだ領域) によりこれらの位置が特定され，それぞれのクラスは *person*，*horse* というラベルにより示されている．またこの画像では，*person* が *horse* に *ride* するという状況が表されている．このとき，主体である *person* を subject，客体である *horse* を object と呼び，これらの間の関係を表す *ride* を predicate と呼ぶ．

オブジェクト間のつながりを表す predicate には，*ride* のような動作を表すもの以外に，*on*，*under* のような位置関係を表すもの，*taller than* のような比較関係を表すものなどが考えられる．従来研究では，動作を扱うもの [2]，主に位置関係に注目するもの [5] というように，predicate の種類を限定するものが多く，前者では人の姿勢等 が有効な手がかりを与え，後者ではオブジェクト間の相対位置座標が有用である [5, 1] とされている．

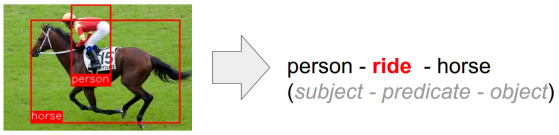

図 1: オブジェクト間の predicate 推定

画像における predicate 推定は，オブジェクトの識別タスクとは性質が異なる．例えば，クラスラベルとして *dog* が与えられるオブジェクトの画像は互いにある程度類似する．一方で，*carry* という predicate が与えられる画像は，その object のクラスによって全く異なる可能性がある．したがって，predicate 推定において考慮すべき画像は多様なバリエーションを有する．

本研究では，特定の種別に限定されない predicate の推定を行うために，言語特徴，画像特徴，領域特徴を統合する．

subject，object の各オブジェクトに対して与えられるクラスラベルの単語から，言語特徴量を得る．word2vec[4] による単語の分散表現を言語特徴量として用いることで，意味的な類似度を捉えることができるので，例えば *elephant* と *giraffe* のような，上位クラス (animal) は共通しているが，画像的には異なるようなオブジェクトに対して，類似した特徴量を与えることができる．

また，画像から bounding box によって切り出された二つのオブジェクト領域が subject と object のどちらであるかを区別したうえで，それぞれの領域に対し，CNN を適用して画像特徴量を得る．例えば，*throw*，*walk on* という動作は，subject として *person* という共通のラベルが与えられるが，その画像的特徴 (姿勢など) は異なるので，オブジェクト領域に対する画像特徴を用いることで，これらの識別が容易になると期待できる．

さらに，*on* や *under* といった位置的関係を表す predicate においては，オブジェクト間の位置関係に注目する必要があるので，それぞれの bounding box の相対位置や面積などを利用して領域特徴量を得る．

これらの言語特徴量，画像特徴量，領域特徴量の三つを同時に扱った predicate 推定を行うニューラルネットワークの構成を提案し，標準的なテストデータ

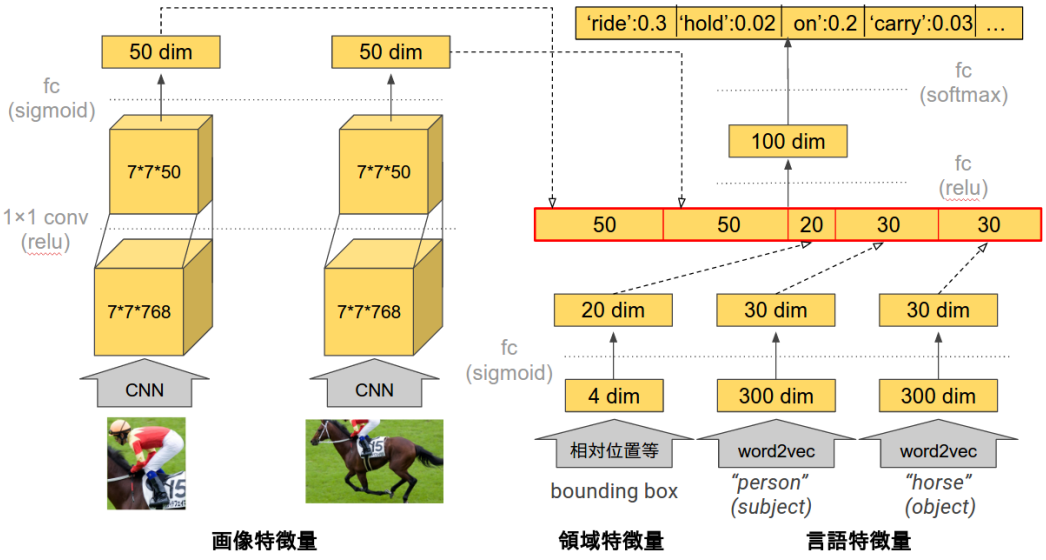

図 2: 各特徴量を統合して predicate 推定を行うニューラルネットワークの構成

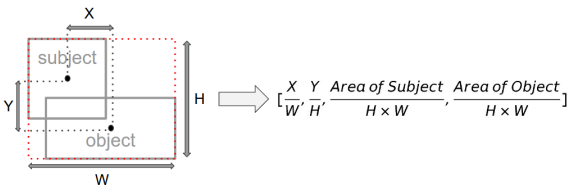

図 3: 領域特徴量の生成

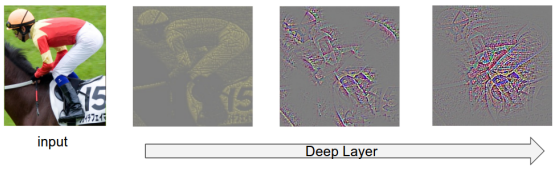

図 4: 畳み込みニューラルネットワークの可視化

セットを用いて評価する．

3 提案手法

言語特徴量，領域特徴量，画像特徴量を入力としたニューラルネットワークの構成を図 2 に示す．

言語特徴量には先ほど述べたように word2vec ベクトルを用いる．具体的には，`subject` と `object` のラベルに対してそれぞれ 300 次元のベクトルを得る．

図 3 に，領域特徴量の生成方法を示す．`subject`，`object` のそれぞれの bounding box から得られる水平距離，垂直距離，`subject` の bounding box の面積，`object` の bounding box の面積という 4 つの情報から，図 3 に示すような正規化を施し，4 次元の特徴ベクトルを領域特徴量として抽出する．

図 5 に，画像特徴量の生成方法を示す．画像特徴量には，CNN の中間層から得た特徴量を用いる．CNN は画像認識などで用いられることが多いニューラルネットワークであり，本研究ではオブジェクトの識別ができるよう学習された VGGnet[6] を使用する．入力画像がこの CNN の各層でのユニットの活性化にどう影響しているかを，[8] を用いて可視化したものを図 4 に示す．

画像特徴量に求められる情報は各オブジェクトのラベルではなく，人の姿勢やオブジェクトの向きなどであるが，オブジェクトの識別を行うネットワークのため，特徴量を抽出する層が深すぎるとその特徴量は識別タスクに特化したものとなってしまう．一方，浅い畳み込み層などを用いると画像におけるエッジ等の特徴が抽出されてしまう．そこで，与えられた画像の全体から bounding box によって切り出された `subject`，`object` のそれぞれの部分画像を CNN に入力し，CNN の畳み込み層の中でちょうど中間の層から得られた 7 × 7 × 256 行列と，最も深い層から得られた 7 × 7 × 512 行列を連結して，7 × 7 × 768 行列の特徴量を抽出する．

7 × 7 × 768 行列とは入力された画像が 7 × 7 のサイズまで畳み込まれ，768 のチャネルをもった特徴量である．ただし実際には bounding box によって切り出された画像は長方形であるため，入力を任意のサイズの出力に変形する手法である ROI Pooling を用いて，畳み込み層から得られた出力を 7 × 7 のサイズの出力に変形する．

これらの言語特徴量，領域特徴量，画像特徴量のベクトルをすべて連結してニューラルネットワークの入力とすると，この入力は非常に高次元のベクトルとなり学習が困難である．そこで，言語特徴量を入力として `predicate` を識別するニューラルネットワーク，領域特徴量を入力として `predicate` を識別するニューラルネットワーク，画像特徴量を入力として `predicate` を識別するニューラルネットワークをそれぞれ構成し

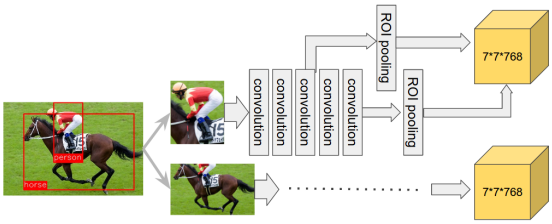

図 5: 画像特徴量の生成

て独立に学習し，これらのニューラルネットワークを連結した図 2 に示すニューラルネットを用いて再度 predicate の識別を学習する．それぞれの特徴量を入力とするニューラルネットワークを図 6 に示す．

最終出力層は，predicate の種類だけユニットを用意し，活性化関数はソフトマックス関数として，1-of-K 表現の教師データを用いる．

表 1: 各特徴量の選択による結果の比較

言語	領域	画像	MAP	識別率
○			0.160	0.522
	○		0.080	0.430
		○	0.078	0.405
○	○		0.190	0.565
○		○	0.168	0.537
	○	○	0.094	0.462
○	○	○	**0.192**	**0.567**

表 2: Lu らの研究との比較

	言語	領域	画像	MAP	識別率
Lu	○		○	**0.295**	0.479
提案手法	○		○	0.168	0.537
提案手法	○	○	○	0.192	**0.567**

4 実験

人手で各オブジェクトに対するラベル，bounding box を評定した 5,000 枚の画像 [3] を用い，4,000 を学習用データとし，1,000 を評価用データとした．学習用データに含まれる subject-object のペアは 30,355 個，評価用データではペアが 7,638 個であった．また，オブジェクトの種類は 100 種，predicate の種類は 70 種であり，70 クラスの識別問題ということになる．

predicate 毎のデータ数のばらつきを図 7 に示す．predicate 毎にそのデータ数は大きく異なり，*on* や *wear* など非常に多い predicate が存在する一方で，*eat* や *skate on* など数個程度の predicate も存在するなど，数のばらつきは非常に大きい．

評価手法としては各画像毎に，画像内の複数の subject-object のペアに対する predicate の識別率を計算し，これを画像の枚数で平均したものを評価値とする．また，画像検索等での評価として，predicate 毎に平均適合率を計算し，これを平均することで求めた MAP (Mean Average Precision) を用いる．

5 結果と考察

表 1 に，言語特徴量，領域特徴量，画像特徴量のそれぞれの有効性を確認するための，各特徴量の組み合わせ 7 パターンで実験を行った結果を示す．ここから，言語情報のみを用いたときの結果と，領域特徴量，画像特徴量，言語特徴量のすべてを用いたときの結果に大差はなく，subject と object のラベルさえ与えられれば，ある程度 predicate を識別することが可能であることが確認できる．また，領域特徴量，画像特徴量，言語特徴量のうち一つのみを用いるより，これらのうち二つの特徴量を用いた場合の方が性能が向上していることがわかる．しかしながら領域特徴量と言語特徴量を用いた場合と，3 つの特徴量すべてを用いた場合ではあまり性能の向上が見られない．この結果は言語特徴量と領域特徴量を合わせたものに対し画像特徴量が新規の情報が含んでいないか，このニューラルネットワークが 3 つの特徴量に含まれる情報を上手く統合できていないかのどちらかであることをを示している．

また，表 2 に画像特徴量，言語特徴量を用いている Lu らの研究 [3] と比較実験を，Lu らと同様の word2vec モデル，CNN を用いて行った結果を示す．ここから，Lu らの研究に比べ，MAP の値が低く，識別率においては性能が向上していることが確認できる．Lu らの研究では predicate 毎に識別器を構成し，それぞれにてランキング学習を用いているため，情報検索などでの評価手法である MAP では性能が高くなったと考えられる．

さらに，表 3 に一部の predicate の識別結果の混同行列を示す．ここから，*near* と *next to* のように明確に区別するのが難しい predicate を識別できていないことが確認できる．画像検索のようなタスクを考えると，*under* をクエリとしたとき *below* が検索対象とならないといった状況はあまり好ましくない．このような問題に対処するために，predicate の類似関係や，包括関係などをいかに定めるかが課題になると考えられる．

表 3: predicate の識別結果の混同行列

識別結果 ＼ 正解	*next to*	*beside*	*near*	*on the right of*
next to	28	7	7	5
beside to	18	4	10	3
near	22	2	9	3
on the right of	15	1	4	13

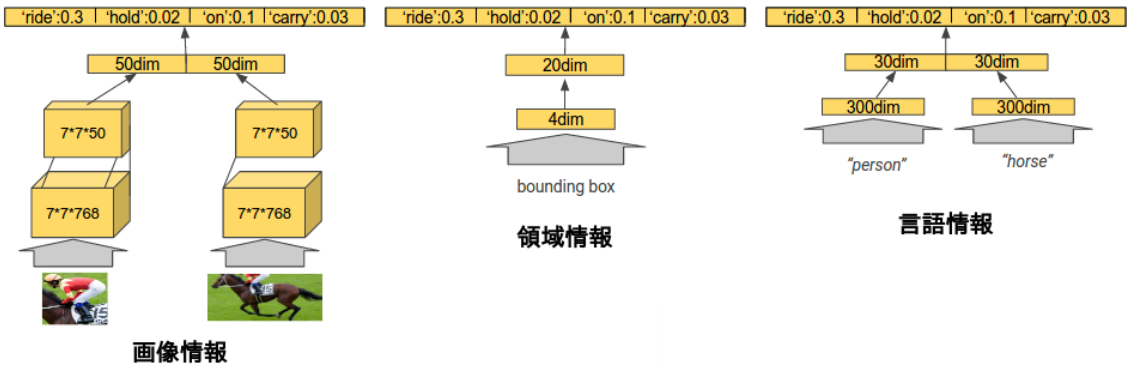

図 6: 各特徴量を個別に用いて predicate 推定を行うニューラルネットワーク

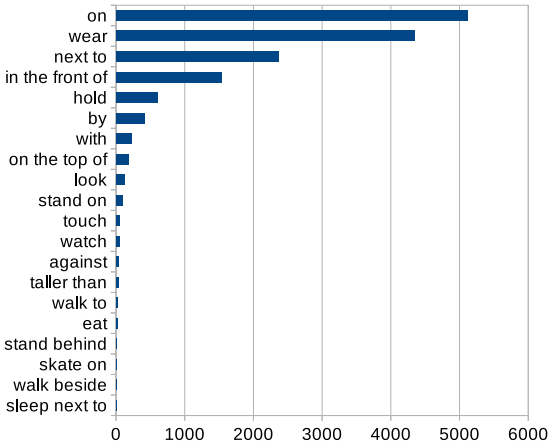

図 7: データセット中の各 predicate のばらつき

6 まとめ

画像内の二つのオブジェクトについて，言語特徴量，領域特徴量，画像特徴量を用いて predicate の識別を行うニューラルネットワークを提案し，MAP，識別率を用いて評価を行った．．各特徴量がそれぞれ predicate 識別に有効であり，それらを二つ組み合わせて扱うことで性能が向上することを示せた一方で，三つの特徴量すべてを統合したことによる性能の向上が見られていない．

今後の課題として，三つの特徴量の統合による性能の向上が見込めるかどうかの調査を行う必要がある．また，画像検索などの応用システムにおいて有用なメタデータを画像に付与するという観点からも，predicate の語彙の体系化などが必要になると考えられる．

参考文献

[1] Desmond Elliott and Arjen P De Vries. Describing Images using Inferred Visual Dependency Representations. *Acl*, pp. 42–52, 2015.

[2] Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 10, pp. 1775–1789, 2009.

[3] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9905 LNCS, pp. 852–869, 2016.

[4] T. Mikolov, J. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.

[5] B. Rosman and S. Ramamoorthy. Learning spatial relationships between objects. *The International Journal of Robotics Research*, Vol. 30, No. 11, pp. 1328–1342, 2011.

[6] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICRL)*, pp. 1–14, 2015.

[7] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image Captioning with Semantic Attention. *Cvpr*, p. 10, 2016.

[8] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8689 LNCS, No. PART 1, pp. 818–833, 2014.