画像によって単語意味表現をエンハンスするニューラルネットワークモデル (ViEW model)

長谷川 美夏　　小林 哲則　　林 良彦
早稲田大学理工学術院
mika@pcl.cs.waseda.ac.jp

1 はじめに

近年，テキストコーパスからは必ずしも得られない・得られにくい実世界に対する知覚情報を言語情報と統合したマルチモーダル意味表現の研究が活発化している [3]．本稿では，画像から抽出した視覚特徴とテキストコーパスから得た分散表現による言語特徴を統合することにより，エンハンスされた単語の意味表現を与えるニューラルネットワークモデル Visually Enhanced Word embedding model (ViEW model) を提案し，単語ペア間の意味的関連度の推定タスクに適用した際の有用性を評価する．

ViEW model のニューラルネットワークは，言語特徴に対する多層のオートエンコーダを基本構造としているが，中間層において，対応する単語の画像データから Convolutional Neural Network (CNN) によって得た視覚特徴と対応付けをするように構成している．これにより，対応する画像を事前に学習できていない単語に対しても，類似の単語から学習した視覚特徴が反映できることを狙っている．

標準的なデータセットである MEN[1] を利用した評価実験によれば，ViEW model によって構成したマルチモーダル意味表現は，言語特徴のみから構成した意味表現よりも意味的関連度推定のタスクにおいて精度が高く，マルチモーダルな特徴を本モデルで統合することの有効性が確認できた．また，視覚特徴が未知の単語についても視覚特徴ベクトルを学習できている単語と同様にマルチモーダル意味表現を構成できることが確認でき，これは本手法により構成する意味表現が広い適用性を持ちうることを示唆する．

2 提案手法：ViEW model

マルチモーダル意味表現は，複数の異なるソースから得た意味表現 (言語, 画像, 音声,etc...) の情報を統合した意味表現であり，言語情報によるユニモーダルな意味表現よりも単語の意味計算において優れていると言われている [3]．人間は概念を単語と結びつける際に，言語情報以外にも視覚・聴覚などから得られる知覚情報を利用しているが，中でも視覚情報は人間が単語の意味を決定するにおいてかなり支配的な情報であり，テキストベースの特徴に関して補足的な情報を提供すると考えられる [1]．これに基づき，本研究でも言語情報に統合する知覚情報として画像データから得た視覚情報を用いることとする．

画像によって単語意味表現をエンハンスするニューラルネットワークとして図 1 に示す 5 層のニューラルネットワーク Visually Enhanced Word embedding model (ViEW model) を提案し，マルチモーダルな特徴の統合を図る．

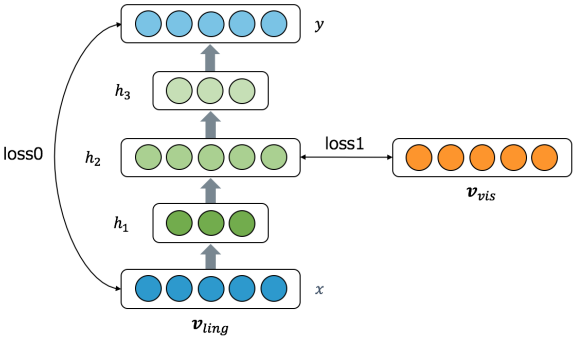

図 1: ニューラルネットワークモデル: ViEW model

ViEW model の基本構造は，各単語の言語特徴ベクトル $\boldsymbol{v}_{ling}$ についてのオートエンコーダであり，$\boldsymbol{v}_{ling}$ の持つ言語情報由来の特徴に，$\boldsymbol{v}_{vis}$ の持つ視覚情報由来の特徴を加えている．これにより，中間層において言語情報に補完的な情報を与える視覚情報が導入され，より人間の感覚に近くなるよう強化された単語の意味表現が構成されることを期待する．

学習は単語の言語による意味表現 $\boldsymbol{v}_{ling}$ が言語による意味表現 $\boldsymbol{v}_{ling}$ として復元されるような Loss (loss0) と，単語の中間表現 h2 を画像による意味表現 $\boldsymbol{v}_{vis}$ に近づけるような Loss (loss1) を設定し，それらが最小化されるように誤差逆伝播を行う．このとき，Loss は最小二乗誤差である．学習に用いるデータは各単語について言語特徴ベクトル $\boldsymbol{v}_{ling}$ と視覚特徴ベクトル $\boldsymbol{v}_{vis}$ の対を用いる．また，本稿の実験では h1，h3 層を 50 次元とした．

マルチモーダル意味表現を導出する際，対象となる単語の言語特徴ベクトルを入力し，中間層 h3 の特徴

ベクトルの値をマルチモーダル意味表現として用いる．入力が視覚特徴ベクトルを学習していない単語の言語特徴ベクトルであった場合，言語特徴ベクトルが類似していれば視覚特徴ベクトルも類似しているという仮定に基づきマルチモーダル意味表現においても，情報が補完できると考えられる．

今回，マルチモーダル意味表現として用いるのは h3 層の特徴ベクトルである．h2 層で言語的特徴に視覚的特徴を含めたあと，h3 層では再び言語的特徴に戻そうとしており，視覚的特徴が与える影響が小さくなる．視覚情報は全てが言語情報に対し効果的であるとは限らないことから，ここで視覚的特徴を抑制することは効果的であると考える．これは特に形容詞など，画像データセット内の画像に一貫性がなく，概念が画像によって表現されにくい単語に有効であると考えられる．

3 特徴ベクトル

ViEW model の学習に用いる特徴量の導出方法を説明する．

3.1 言語特徴ベクトル

言語特徴ベクトルは英語版 Wikipedia の大規模コーパスから Mikolov らの提案する Word2Vec[7] の分散表現を用いて導出する．

言語特徴ベクトルの導出にあたって英語版 Wikipedia のコーパスデータである enwik9[1]を用いる．これは 2006 年 3 月 3 日の英語版 Wikipedia の dump データの先頭から 10^9bytes 分のコーパスデータを用いており，英語で自然言語処理をする研究では頻繁に用いられている．

Word2Vec の Skip-gram モデルを用いて文脈窓幅 5，次元数 300 のパラメータで enwik9 を学習した言語特徴ベクトルを導出し，マルチモーダル意味表現を構成するための ViEW model の学習データとして用いる．

3.2 視覚特徴ベクトル

視覚特徴ベクトルは大規模画像データセット ESP-Game dataset[9][2]，ImageNet[5][3]の画像から GoogLeNet[8] を用いて導出している．

画像から視覚特徴ベクトルを抽出する方法の代表的なものとして bag-of-visual-words (BoVW) と CNN 中間層特徴量が挙げられる．歴史的には BoVW が用いられているものが多いが，近年は画像分野のタスクにおいて CNN の精度が注目されている．本研究に用いた GoogLeNet は，画像認識技術のコンテストである ILSVRC2014 において首位の成績を収めた 22 層のディープニューラルネットワークである．ImageNet で学習済みの chainer GoogLeNet モデル[4]の pool5/7x7_s1 層を利用して各画像について 1024 次元の視覚特徴ベクトルを導出した．pool5/7x7_s1 層は最終層の手前の Pooling 層である．GoogLeNet は最終層で ImageNet の synset に基づく 1000 カテゴリに分類されるため，その手前の層には物体の特徴が表れていると考えられる．

用いる 2 種類のデータセットについて述べる．

- ESP-Game dataset： the ESP game と呼ばれるシステムにより収集された画像とラベルのセット．収録される画像の特徴として対象オブジェクトが画像の中心にあるとは限らないことや，画像 1 枚に複数の単語がラベル付けがされているという特徴が挙げられる．ラベルは意味的関連度の評価時の単語の網羅性を上げるため，nltk の WordNetLemmatizer を用いて見出語化して利用する．
- ImageNet： 2016 年現在，約 2 万 2 千の synset (語彙概念) に約 1400 万枚の画像が割り当てられているデータセットである．画像の中心に対象オブジェクトがある場合が多いことや，synset をノードとして概念が階層化されているという特徴が挙げられる．

本研究では，各単語に対する視覚特徴ベクトルが必要であるため ESP-Game dataset，ImageNet はそれぞれデータセット内に画像が 50 枚以上存在する単語を用いる．1 つの単語 (ラベル) に 100 枚以上画像が存在している場合はランダムに 100 枚のみを利用し，複数のラベルが付与されている画像については別ラベルにおいて画像の重複を許す．それらの画像について GoogLeNet の中間層 (pool5/7x7_s1) から 1024 次元の特徴量を抽出する．

前述の特徴量を用いて，各単語について以下の 3 種類の視覚特徴ベクトルを構成し比較する．ニューラルネットワークモデルで学習をする際に入力に用いた言語特徴ベクトルに対し，視覚特徴ベクトルの次元が大きいと中間層 h2 のユニット数が大きくなるため，PCAn，AEn ではそれぞれ主成分分析，オートエンコーダを用いて次元削減を行う．

- Mean：各単語について 50〜100 枚の画像から得た特徴量を各次元ごとに平均した 1024 次元の特徴ベクトル．
- PCAn：各単語について導出した Mean の 1024 次元特徴ベクトルを主成分分析 (PCA) により次元数が言語特徴ベクトルの次元数 n と等しくなるように次元削減した特徴ベクトル．
- AEn：各単語について導出した Mean の 1024 次元特徴ベクトルを 5 層の AutoEncoder により次元数が言語特徴ベクトルの次元数 n と等しくなるように次元削減した特徴ベクトル．2 層目 4 層目

[1] http://mattmahoney.net/dc/textdata.html
[2] http://hunch.net/~jl/
[3] http://image-net.org
[4] https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet

の次元数を 2048 次元，3 層目の次元数を n とし，3 層目から得た中間特徴量を用いる．

4 評価手法

4.1 単語間の意味的関連度

意味的関連度 (Semantic relatedness) [4, 2] とは，単語間の意味的な関連の程度を示す指標である．自然言語処理において意味的関連度の導出には膨大な量のテキストデータが必要であるとされる．人間は概念の意味的関連度を判断するとき，テキスト情報だけではなく知識や経験も踏まえて判断するため，その差を埋めるのが課題となる．また，意味的関連度は意味的類似度 (Semantic Similarity) よりも広義な概念であり，本稿では両者は区別して用いる．意味的関連度には car/wheel のような全体-部分の関係，hot/cold といった対義性の関係や pencil/paper のような機能を表す関係も含まれる．

単語間の意味的関連度を定量化する手法として，ベクトル化した単語のコサイン類似度を計算するものが挙げられる．このコサイン類似度の値が大きいほど意味的関連度が大きいとみなせる．

4.2 評価データセット：MEN

評価データには，人手によってスコア付けされた単語ペアの類似度データセットが用いられる．本研究で用いる MEN[1] は，3000 ペア，751 単語について単語ペアの意味的関連度を人手によりスコアリングされたデータセットである．意味的関連度のスコアの範囲は 0-1 であり，例えば，意味的関連度が高いものは beach/sand が 0.96, 意味的関連度が低いものは bakery/zebra が 0 である．また，単語にはそれぞれ品詞が指定されている．

4.3 評価指標

評価指標は，Spearman の順位相関係数を用いる．MEN の単語ペア間について，マルチモーダル意味表現のコサイン類似度の系列と MEN のスコアの系列について Spearman の順位相関係数を求め，この値が大きいほど意味表現が人間の意味的関連度の評価スコアに近く優れた意味表現であるとする．

5 評価実験

5.1 実験：マルチモーダル意味表現の構成

視覚特徴ベクトルと統合するためのベースラインの言語特徴ベクトルとして enwik9 をコーパスとした文脈窓幅 5，300 次元のベクトル (表 1 中，ew9.300) を用いる．このパラメータは予備実験により，文脈窓幅と次元数を変更したところ最も MEN との相関係数が高かったものである．この言語特徴ベクトルについて，前述した 2 通りの画像データセットについて 3 通りの視覚特徴ベクトルを用いてマルチモーダル意味表現を評価する．

表 1: マルチモーダル意味表現と MEN のスコアの相関係数

ESP-Game				
		Ling.	Visual	Multimodal
ew9.300	全体	0.74	-	0.74
(Mean)	画像あり	0.75	0.58	0.75
ew9.300	全体	0.74	-	**0.76**
(PCA300)	画像あり	0.75	0.54	0.78
ew9.300	全体	0.74	-	0.74
(AE300)	画像あり	0.75	0.59	0.74
ImageNet				
		Ling.	Visual	Multimodal
ew9.300	全体	0.74	-	0.75
(Mean)	画像あり	0.74	0.59	0.77
ew9.300	全体	0.74	-	**0.77**
(PCA300)	画像あり	0.74	0.67	0.8
ew9.300	全体	0.74	-	0.75
(AE300)	画像あり	0.74	0.59	0.77

表中の Ling. は言語特徴ベクトルのみの相関係数，Visual は視覚特徴ベクトルのみの相関係数，Multimodal はマルチモーダル意味表現の相関係数を示す．画像ありはデータセットに画像が存在していた単語のみの結果である．

表 1 より，視覚特徴ベクトルは中間層の特徴量をそのまま用いるよりも PCA によって次元削減をした方が，より言語特徴ベクトルに対して補完的に働くと考えられる．一方で，AE300 の結果は Mean の結果とほぼ同様のスコア推移であることから，次元削減すること自体ではなく PCA による次元削減が効果的であると言える．これらの結果を見ると，ESP-Game よりも ImageNet の方が高スコアの傾向があるがそれらの差は僅かである．ESP-Game は複数タグが付与されているため画像内の物体の共起関係が取れる，一方で ImageNet は画像内の物体はラベルが示すもののみであるという特性を持っているが，どちらが画像特徴量を抽出するデータセットとして適切であるかは断定できない．ベースラインの言語特徴ベクトル (300 次元) よりも小さい次元数 (50 次元) で高いスコアが得られていることは，単語の特徴ベクトルを用いた計算を行う場合の計算コストの低下という効果も期待できる．

5.2 Lazaridou らの関連研究 [6] との比較

Lazaridou らの研究では，enwik9 から Skip-gram モデルによる 300 次元の言語特徴ベクトルを構成し，ImageNet から CNN (CaffeNet) を用いて 4096 次元の視

覚特徴ベクトルを抽出し，それらを用いてマルチモーダル意味表現を構成している．彼らは，Skip-gram の式を拡張し言語表現と視覚表現のコサイン類似度を高めるように埋め込みを行う Multi-modal Skip-gram というモデルを提案している．彼らの実験結果は，視覚情報が存在するものだけで行った評価と，視覚情報を補完した場合の評価がほぼ同一であることから，視覚情報がないものについても，十分利用可能なマルチモーダルな意味表現が構成できていると述べている．このモデルは視覚情報を単語に伝播するため，学習データのない画像のラベリングや検索ができることも述べられている．

表 2: Lazaridou らの研究との実験結果の比較

ESP-Game				
		Ling.	Visual	Multimodal
ViEW	全体	0.74	-	**0.76**
(PCA300)	画像あり	0.75	0.54	**0.78**
ImageNet				
		Ling.	Visual	Multimodal
ViEW	全体	0.74	-	**0.77**
(PCA300)	画像あり	0.74	0.67	**0.8**
Lazaridou	MEN(100%)	-	-	0.75
MMSG-A	MEN(42%)	-	-	0.74
Lazaridou	MEN(100%)	-	-	0.74
MMSG-B	MEN(42%)	-	-	0.76

Lazaridou らは Ling.，Visual のそれぞれの結果を提示していなかったため，相関係数の向上率は不明である．ViEW の結果は，Lazaridou らの結果よりも ViEW model の h3 層から構成したマルチモーダル意味表現が全て相関係数が高いスコアを示している．また，Lazaridou らは全体を評価した時と視覚特徴が存在するもののみ評価する場合でスコアがほぼ変わらないことから補完ができていると示しているが，ViEW においては全体を評価した時の方が僅かにスコアが下がっている．しかし，ViEW model の結果の方がスコアが良いため，期待する結果が得られたと言える．

6 おわりに

画像によって単語意味表現をエンハンスするニューラ ルネットワークモデル Visually Enhanced Word embedding model (ViEW model) を提案し，その中間層から取り出した特徴量をもとに各単語のマルチモーダル意味表現を構成した．

実験では，言語特徴ベクトル (ベースライン) と比較すると，ViEW model によって構成した視覚特徴ベクトルを統合したマルチモーダル意味表現は全て意味的関連度の推定において優れた結果となった．中でも，PCA による次元削減時のマルチモーダル意味表現が優れている傾向にあった．性質の違う 2 種類の視覚情報源のデータセットについては，同程度のスコアを得たためどちらが視覚情報源として今回のタスクに適していると断言することはできなかった．

関連研究との比較の結果，意味的関連度の推定タスクにおいて本論文で提案した ViEW model から得たマルチモーダル意味表現は同様もしくは若干優れた結果を得られた．以上より，ViEW model により構成したマルチモーダル意味表現は意味的関連度の推定タスクにおいて有効に働くことが確認できた．

今後は，品詞や具体度といった単語の特性を考慮すること，意味的関連度以外にクロスモーダル検索といった別のタスクにおいてどの程度効果を得られるのか調査に取り組む．

参考文献

[1] Elia Bruni, Daniel Gatica-perez, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.*, Vol. 49, No. December, pp. 1–47, 2014.

[2] Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Comput. Linguist.*, Vol. 32, No. 1, pp. 13–47, 2006.

[3] Yansong Feng and Mirella Lapata. Visual Information in Semantic Representation. *Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter ACL, Los Angeles, California, June 2010*, Vol. 9, No. June, pp. 91–99, 2010.

[4] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 1606–1611, 2007.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[6] Angeliki Lazaridou, The Pham Nghia, and Marco Baroni. Combining Language and Vision with a Multimodal Skip-gram Model. *Proc. Hum. Lang. Technol. 2015 Annu. Conf. North Am. Chapter ACL, Denver, Color. May 31 June 5, 2015*, pp. 153–163, 2015.

[7] Tomas Mikolov, I. Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Proc. NIPS*, Vol. 9, pp. 1–9, 2013.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Vol. 07-12-June, pp. 1–9, 2015.

[9] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. *Proc. 2004 Conf. Hum. factors Comput. Syst. - CHI '04*, pp. 319–326, 2004.