画像認識器の物体ラベルを活用した単語の特徴表現

村岡雅康　　那須川哲哉
mmuraoka@jp.ibm.com　nasukawa@jp.ibm.com
日本アイ・ビー・エム株式会社 東京基礎研究所

1 はじめに

自然言語処理で画像情報を利用する取り組みにおいては，これまで，畳み込みニューラルネットワーク (CNN) の中間層がよく利用されてきた．画像処理分野において高精度かつ多種多様な物体の認識が可能となり，その技術を画像認識サービスとして提供する企業やベンダーが増えてきている中，そのサービスで提供される結果をそのまま単語の特徴表現に使用することができれば，画像処理分野の進化を自然言語処理に取り込むことが容易になる．そこで本稿では，画像認識サービスから得られる物体ラベルを，単語の特徴表現として使用した場合の有用性について調査する．

画像認識サービスの特長の一つとして，認識できる物体の種類数が，既存の CNN よりはるかに多いことが挙げられる．一例として，IBM が提供している IBM Watson Visual Recognition[1] (以降，Watson VR と呼称) のユニークな物体ラベルの種類は，2018 年 9 月時点で 1 万 3 千種類を超えていた．これに対し，既存の CNN モデルで予測できる物体数は高々1000 クラスである．これは，既存のモデルが，ILSVRC[15] と呼ばれる画像認識タスク用に開発されたデータセットを用いて学習されていることに由来する．既存のモデルを使えば，任意の層から特徴量を取り出すことが可能だが，画像認識サービスが提供するのは，基本的に物体ラベルとその確信度，すなわち，CNN モデルに対応する最終層のみである．これまで，CNN の中間層が他のタスクで有益であると経験的に示されてきたため，最終層は捨てられてきた [5, 8, 19]．しかし，最終層も他のタスクに活用できることが示せた場合，そのようなサービスの研究としての利用価値が今後さらに高まるだろう．

本稿では，画像認識器の出力である物体ラベルを用いて構築した単語の特徴表現を Lexical Entailment (LE) タスクによって評価する．LE とは 2 つの概念間に上位–下位の関係が成り立つ場合を指し，「X は Y の一種である」の形で表現できる．この時，Y は X の上位概念であるという．LE の認識は文間の含意関係認識 [2, 4] や，オントロジーの自動構築 [11, 18]，メタファー検出 [10] に必要となる重要な技術の一つである．評価実験では，次の 2 種類の能力を評価する：2 つの単語が与えられた時，(1) それらが LE 関係にあるかどうかを検出する能力 (LE 検出能力)，および，(2) LE 関係にある場合，どちらが上位概念かを特定する能力 (向き識別能力) である．

2 関連研究

2.1 DIH に基づく手法

Rimell[14] および Santus ら [16] は，下位概念が出現する文脈は，上位概念が出現する文脈より情報量があるという仮説 (Distributional Informativeness Hypothesis, DIH) を立てた．この仮説に基づき，Santus ら [16] は単語の特徴表現を，その語と共起しやすい単語群によって構築した．本研究では，単語と共起しやすい単語の代わりに，ある単語に関連する画像に現れる物体[2]を用いることで，DIH に基づく手法を提案する．

2.2 画像特徴量を用いた単語の特徴表現

Kiela ら [8] は，学習済み CNN から得られた画像の特徴量を用いて単語の特徴表現を構築した．具体的には，画像検索エンジンを用いて単語に関連する画像を取得し，それらに画像認識用 CNN[7] を適用することで画像の特徴量を得る．また，上位–下位概念の向きの識別を行うために，Deselaers および Ferrari[3] が示した，“上位概念に関連する画像のばらつきは，下位概念に関連する画像のばらつきよりも大きい” という性質も用いている．これは直感的には，図 1 に示すように，「animal」という語を画像検索すると，種類の異なる動物の画像が取得されるのに対して，その下位概念である「tiger」の画像検索結果はほぼ全てトラの画像になっていることから理解できる．

[1] https://www.ibm.com/watson/services/visual-recognition/

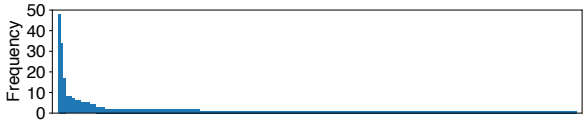
[2] 正確には，画像認識サービスによって認識された物体ラベル．

(a) 「animal」の画像検索結果

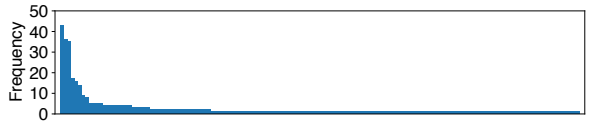
(b) 「carnivore」の画像検索結果

(c) 「tiger」の画像検索結果

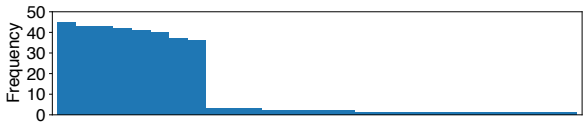
図 1: Google 画像検索エンジンの検索結果.



(a) animal (ラベルの種類数: 175)



(b) carnivore (ラベルの種類数: 145)



(c) tiger (ラベルの種類数: 28)

図 2: 物体ラベルのヒストグラム (降順).

3 提案手法

3.1 物体ラベルを用いた単語の特徴表現

まず，単語に関連する画像を，Kiela ら [8] と同様に，Google 画像検索[3]を使用して取得する．ここで，1 単語あたりの取得画像数を l とする．本稿では，$l = 50$ とした．図 1 に画像検索結果例を示す．animal，carnivore，tiger の順に画像中の物体のばらつきが小さくなっていることが確認できる．

続いて，得られた各画像に対して，Watson VR を適用し，$m = 10$ 個の物体ラベルとそれぞれのラベルの確信度を得る．図 2 は，図 1 の結果に Watson VR を適用した結果の物体ラベルヒストグラムである．ラベルは頻度の高い順に列挙している．図から，より意味の広い語ほど long-tail となっており，ラベルの種類数が多いことがわかる．

以上の処理から，各単語に lm 個の物体ラベルが付与される．データ全体のラベルの異なり数を N_L とすると，単語の特徴表現は $l \times N_L$ 次元の行列 $\boldsymbol{V}$，もしくは，集約した N_L 次元のベクトル $\vec{v}$ で表される．表現形式は，次節で述べる意味的広さを求める関数によって使い分ける．また，集約方法については，列ごとの，平均 (*avg*)，最大値 (*max*)，平均と分散の結合 (*st*)[5] の 3 種類を試し，実験において最適な集約方法を求める．

3.2 意味的広さの定量化

LE タスクを解く上で単語の意味の広さを定量化する必要がある．本研究では，DIH に基づいて提案された平均情報量 [16, 17] を用いる：

$$\mathtt{ent}(\vec{v}) = -\Sigma_{i=1}^{N_L} p(v_i) \log_2 p(v_i). \quad (1)$$

ただし，v_i は $\vec{v}$ の i 番目の要素であり，$p(v_i) = v_i / \Sigma_j \vec{v}_j$ である．

比較のために，画像の特徴量を用いて構築した単語の特徴表現とともに使用された [8]，以下の 2 種類の関数も実験において検証する．

$$\mathtt{ca}(\boldsymbol{V}) = \frac{2}{l(l-1)} \Sigma_{i<j\leq l} \{1 - \cos(\vec{v_i}, \vec{v_j})\}, \quad (2)$$

$$\mathtt{cc}(\boldsymbol{V}) = \frac{1}{l} \Sigma_1^l \{1 - \cos(\vec{v_i}, \vec{\mu})\}. \quad (3)$$

$\vec{v_i}, \vec{v_j}$ はそれぞれ $\boldsymbol{V}$ の i, j 番目の行ベクトルであり，$\vec{\mu} = \Sigma_i \vec{v_i} / l$ は $\boldsymbol{V}$ の重心である．

3.3 LE 検出と向き識別

2 つの単語 x, y が与えられた時，それらが LE の関係にあることを検出する関数を $det(x, y)$ で表し，$det(x, y) \geq \alpha_{det}$ の時，x, y は上位–下位の関係にあるとする．α_{det} は閾値である．関数 $det(x, y)$ として，本研究では，コサイン類似度 (*cos*) と KL ダイバージェンスに対称性を持たせた JS ダイバージェンス (*js*)[9] を用いる．Kiela ら [8] はコサイン類似度を使用しているが，これは数値の大きな共通要素があると類似度も高くなってしまうため，ベクトル全体が似ているときに類似度が高くなる JS ダイバージェンスも試す．

また，x と y のどちらが上位概念であるかを識別する関数 $dir(x, y)$ は以下で表される [8, 16].

$$dir(x, y) = 1 - \frac{f(\vec{v}_x)}{f(\vec{v}_y)}. \quad (4)$$

[3] https://images.google.com/

表 1: LE タスクにおける正解率 (%)．カッコ内は交差検定時の標準偏差を表し，最適パラメータは意味的広さ，特徴表現の値，集約方法，LE 検出関数の順に列挙．

データセット	BLESS(向き識別)		WBLESS(LE 検出)		BIBLESS(LE 検出&向き識別)	
手法	正解率	最適パラメータ	正解率	最適パラメータ	正解率	最適パラメータ
提案手法	**93.49**	ent, *score, avg*	56.10 (0.04)	ent, *score, avg, cos*	**60.85** (0.03)	ent, *score, avg, js*
CNN(中間層)[8]	91.17	ent, *freq, max*	**57.33** (0.04)	ent, *freq, max, cos*	60.61 (0.03)	ent, *freq, max, cos*
CNN(最終層)[8]	88.56	cc, *score, st*	54.19 (0.04)	cc, *score, avg, cos*	57.97 (0.03)	ent, *freq, max, cos*
DIH モデル [16]	54.97	ent, *score*	52.62 (0.03)	ent, *score, cos*	48.11 (0.02)	ent, *score, cos*
単語埋め込み [1]	65.59	ent, *score*	50.76 (0.02)	ent, *score, cos*	51.30 (0.02)	ent, *freq, js*

ただし，$\vec{v}_x, \vec{v}_y$ はそれぞれ単語 x, y の特徴表現であり，$f(\cdot)$ は 3.2 節で導入した意味的広がりを求める関数である．$dir(x, y) > \alpha_{dir}$ の時，y は x の上位概念とする．ただし，α_{dir} は閾値である．

閾値 $\alpha_{det}, \alpha_{dir}$ は後述の交差検定によって決定する．

4 実験

4.1 実験設定

評価データ　本稿では，提案手法の有用性を LE タスクにおいて評価する．評価データは，LE 検出能力と向き識別能力を測るために Kiela ら [8] によって作成されたものを使用する．BLESS，WBLESS，BIBLESS の 3 種類のデータセットからなり，全てのインスタンスは (x, y, LE ラベル) の 3 つ組で構成される．BLESS は向き識別，WBLESS は LE 検出，BIBLESS はその両方の能力を同時に測ることができる．

ハイパーパラメータ　特徴表現の構築には以下に示すハイパーパラメータが存在する．実験では，以下の全ての組み合わせをグリッド探索によって検証し，各データセットにおいて最も性能の良かったパラメータを報告する．

- 特徴表現の値：出現頻度 *freq*，確信度 *score*
- $\boldsymbol{V}$ から $\vec{v}$ への集約方法：*avg*，*max*，*st*
- 意味的広さ：ent，ca，cc
- **LE** 検出関数：*cos*，*js*

閾値　閾値 $\alpha_{det}, \alpha_{dir}$ はデータセットごとに次のように設定した．BLESS データでは全てのインスタンスにおいて常に y が上位語であるため，$\alpha_{dir} = 0$ とする．また，WBLESS 及び BIBLESS データでは，既存研究 [12, 20] と同様に，データセットから 2%のインスタンスをランダムにサンプリングし，残りの 98%で評価を行うことで閾値を決定する．最終的な正解率はこれを 1000 回繰り返した平均値を報告する．

比較手法　比較手法として，(1) CNN モデルを用いた特徴表現 [8]，(2) DIH に基づく特徴表現 [16]，(3) 単語埋め込みの 3 種類と比較する．CNN のモデルとして 161 層の DenseNet[6] を使用し，中間層 (全結合層の最終層，2,208 次元)，および，最終層 (1,000 次元) から特徴表現を構築した[4]．また，DIH に基づく特徴表現を構築するために，約 9 千万語からなる Reuters corpus (RCV1)[13][5]を使用した．単語埋め込みは 6 千億語のコーパスで事前学習された fastText [1][6]を用いた．

4.2 結果

結果を表 1 に示す．BLESS，BIBLESS データにおいて，既存手法を上回った．WBLESS データにおいても，提案手法は CNN の中間層に迫る結果であると言える．CNN の最終層は 1000 クラスの物体ラベルの予測分布であるが，提案手法との比較から，画像認識器が認識できる物体数が多いほど，LE タスクにおいて効果的な特徴表現を構築できると予想される．

また，画像情報を用いる提案手法及び CNN と，テキスト情報を用いる DIH モデルと単語埋め込みには性能に大きな差が認められる．これは次のような原因が考えられる．DIH に基づく手法 [16] は特徴表現を構築する際，単語の頻度情報を用いるため，使用するコーパスサイズに大きく依存するが，提案手法はこれに依存しない．また，単語埋め込みの学習では，上位–下位概念のように意味の広さを考慮できていない．

最適なパラメータに注目すると，CNN の最終層以外では意味的広さを定量化するのに平均情報量 (ent) が適しており，CNN の中間層を用いた場合，出現頻度 (*freq*) を使用した方が効果的な特徴表現が得られることがわかる．また，集約方法には手法との相性がある

[4] 1 画像あたり全ての特徴量を使うより，値の上位 10 個 (提案手法と同じ数) を使う方が，一貫して性能が良いことが確認されたため，実験では上位 10 個を使用した場合の結果を報告する．

[5] SpaCy (https://spacy.io) を用いて前処理を行い，特徴表現の構築には名詞，動詞，形容詞のみを用いた．

[6] https://fasttext.cc/docs/en/english-vectors.html

ことがわかった．さらに，期待に反して，JS ダイバージェンス (*js*) よりコサイン類似度 (*cos*) が，概して良い結果をもたらすことにも注目したい．

5 結論

本研究では，既製の画像認識器の物体ラベルを活用した手法を提案し，上位–下位概念を識別・検出する Lexical Entailment タスクにおいてその有用性を示した．評価実験では，2 種類のデータセットにおいて既存手法を上回ることを確認した．本手法の特徴表現は言語に依存しない構築のため，言語が異なる単語ペアの比較にも適用可能である．例えば，那須川ら [21] が提案した，画像とテキストが紐付いたデータから様々な認識対象の名称を獲得するタスクに適用すれば，未知の言語における上位–下位概念の獲得につながると考えられる．

謝辞　本稿の取り組みに関し，画像認識器の活用で IBM Research – Thomas J. Watson Research Center の Bishwaranjan Bhattacharjee 氏の，また，多くの議論を通して日本アイ・ビー・エム株式会社 東京基礎研究所（当時）の Khan Md. Anwarus Salam 氏の多大なご支援，ご助言をいただきました．ここに記して感謝致します．

参考文献

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *TACL*, Vol. 5, pp. 135–146, 2017.

[2] I. Dagan, D. Roth, M. Sammons, and F. M. Zanzotto. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. 2013.

[3] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *CVPR*, pp. 1777–1784, 2011.

[4] D. Garrette, K. Erk, and R. Mooney. Integrating logical representations with probabilistic information using markov logic. In *IWCS*, pp. 105–114, 2011.

[5] J. Hewitt, D. Ippolito, B. Callahan, R. Kriz, D. T. Wijaya, and C. Callison-Burch. Learning translations via images with a massively multilingual image dataset. In *ACL (Vol. 1)*, pp. 2566–2576, 2018.

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pp. 2261–2269, 2017.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pp. 675–678, 2014.

[8] D. Kiela, L. Rimell, I. Vulić, and S. Clark. Exploiting image generality for lexical entailment detection. In *ACL-IJCNLP (Vol. 2)*, pp. 119–124, 2015.

[9] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, Vol. 37, No. 1, pp. 145–151, 1991.

[10] M. Mohler, D. Bracewell, M. Tomlinson, and D. Hinote. Semantic signatures for example-based linguistic metaphor detection. In *Metaphor in NLP*, pp. 27–35, 2013.

[11] R. Navigli, P. Velardi, and S. Faralli. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI (Vol. 3)*, pp. 1872–1877, 2011.

[12] K. A. Nguyen, M. Köper, S. Schulte im Walde, and N. T. Vu. Hierarchical embeddings for hypernymy detection and directionality. In *EMNLP*, pp. 233–243, 2017.

[13] N. I. of Standards and T. (U.S.). Reuters corpora, 2018.

[14] L. Rimell. Distributional lexical entailment by topic coherence. In *EACL*, pp. 511–519, 2014.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, Vol. 115, No. 3, pp. 211–252, 2015.

[16] E. Santus, A. Lenci, Q. Lu, and S. Schulte im Walde. Chasing hypernyms in vector spaces with entropy. In *EACL (Vol. 2)*, pp. 38–42, 2014.

[17] V. Shwartz, E. Santus, and D. Schlechtweg. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *EMNLP (Vol. 1)*, pp. 65–75, 2017.

[18] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Coling-ACL*, pp. 801–808, 2006.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pp. 3156–3164, 2015.

[20] I. Vulić and N. Mrkšić. Specialising word vectors for lexical entailment. In *NAACL (Vol. 1)*, pp. 1134–1145, 2018.

[21] 那須川哲哉, 村岡雅康. 君の名は －画像認識対象の名称獲得－. In *NLP*, 2019.