ゼロショット物体認識における辞書定義文の援用

菊池 康太郎 †　　林 良彦　　小林 哲則
早稲田大学理工学術院
† kotaro@pcl.cs.waseda.ac.jp

1　はじめに

人が認識できるほど多種多様な物体を認識するシステムを構築する上で障壁となるのは，学習用画像の収集である．ウェブで学習用画像を収集する場合，物体とその物体が描写された画像の枚数はロングテールの関係になることが知られている [9]．対象の物体が描写された画像が少量である場合，もしくは取得できない場合には，大量の画像の学習を必要とする物体認識システムは適さない．このような問題を扱うタスクの一つがゼロショット物体認識である．ゼロショット物体認識では学習事例のない物体の認識を目的とする．認識システムの学習時に画像が与えられるクラスを既知クラス，画像が与えられないクラスを未知クラス，クラスを説明する非視覚情報を補助情報と呼ぶ．未知クラスの補助情報が事前に取得できるという仮定のもと，既知クラスの画像と補助情報の対応関係から未知クラスの画像と補助情報の対応関係を推定し，画像に尤もらしいクラスを割り当てることで未知クラスの物体認識を実現する．

本研究では補助情報に辞書定義文を用いることを提案する．辞書定義文は収集が容易であることに加え，人がある概念と他の概念を区別するのに必要な情報を持つ．辞書定義文から特徴抽出する４つの手法を提案し，大規模画像データセットを用いて実験を行なった．その結果，提案手法の全てが既存手法のゼロショット認識性能を上回ることを確認し，辞書定義文が補助情報として有効に活用できることを示した．また，４つの手法のうち，パラメータの少ない単純な手法が過学習を抑えられ，良い性能を示すことを確認した．

2　関連研究

大規模な認識問題への適応を目的として，物体認識とは異なる目的で作成されたウェブ資源を補助情報として活用する研究が提案されている．例えば，構造化データである Linked Open Data（LOD）や非構造化データである Wikipedia といった知識資源を活用するものがある [4, 2]．

中でも特に多くの研究で使われる補助情報が単語埋め込み（単語の分散表現）である [3, 7, 1, 8]．代表的な単語埋め込みは，同じ文脈で現れる単語は似た意味を持つ傾向にあるという分布仮説に基づき，大規模なテキストコーパスを使った教師なし学習によって得られる [6] ．公開されている学習済みの単語埋め込みがゼロショット物体認識にそのまま転用できることに加え，ドメインに応じて収集した大量の文からカスタマイズされた単語埋め込みを作成することもできるため，単語埋め込みは実用的な補助情報と考えられている．一例として，テキストコーパスと単語埋め込みの学習方式を詳細に設計することで，ゼロショット物体認識の性能が改善したという報告もある [1]．

3　手法

多くの研究ではクラス名の単語埋め込みをそのクラスを表す特徴表現として使用するのに対し，本研究ではクラス名の単語埋め込みと辞書定義文から得た特徴表現を合わせて使用する．クラス名の単語埋め込みは，その単語がどのような文脈と共起したかという情報を持ち，辞書定義文から得た特徴表現は，その単語が異なる単語を用いてどのように説明できるかという情報を持つ．クラス名の単語埋め込みに加えて，異なる情報を持つと予想される辞書定義文から得た特徴表現を用いることで，画像との対応関係の学習に有効な手掛かりを増やすことが期待できる．

DEm[8] は複数の補助情報を統合して扱うことができるが，辞書定義文の利用を想定した特徴抽出器の設計は行われていない．提案手法は DEm をベースとし，辞書定義文の利用を想定して設計された４種類の特徴抽出器を持つ．提案手法の全体像を図１に示す．

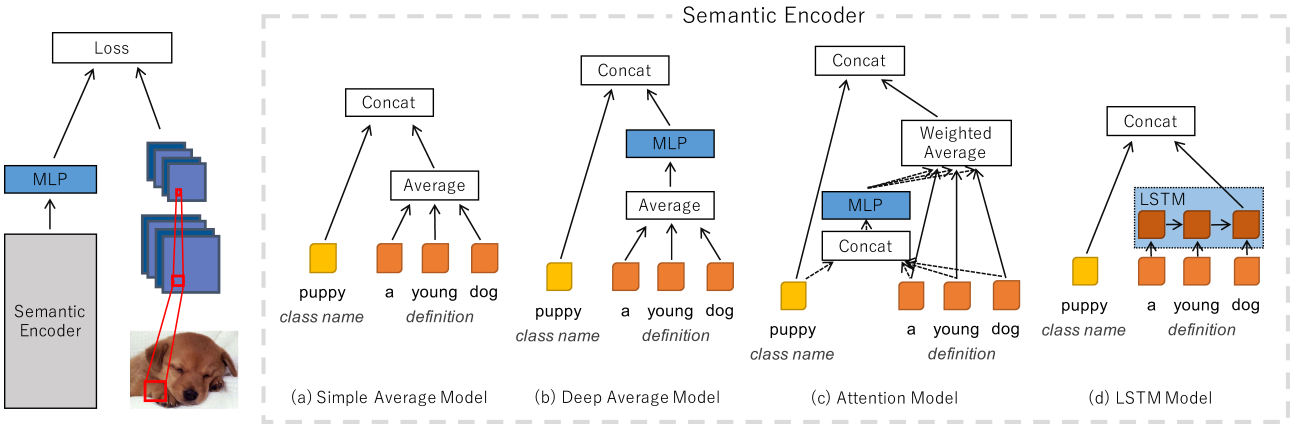

図 1: 提案手法の全体像

3.1 問題設定

画像 $\boldsymbol{I}_i \in \mathcal{I}_{\mathrm{tr}}$ とそのクラスラベル $t_i \in \mathcal{T}_{\mathrm{tr}}$，そしてクラスラベルから得られる補助情報 $\mathcal{Y}_{t_i}$ の組からなる学習用データセット $\mathcal{D}_{\mathrm{tr}} = \{(\boldsymbol{I}_i, t_i, \mathcal{Y}_{t_i}) | 0 \leq i \leq N\}$ を考える．ゼロショット物体認識の目的は，学習用データセットには存在しないラベル $t_j \in \mathcal{T}_{\mathrm{te}}$ $(\mathcal{T}_{\mathrm{tr}} \cap \mathcal{T}_{\mathrm{te}} = \emptyset)$ が付与されたテスト画像 $\boldsymbol{I}_j$ のラベルを補助情報を用いて推定することである．ただし，任意のクラスラベル $t_n \in \mathcal{T}_{\mathrm{tr}} \cup \mathcal{T}_{\mathrm{te}}$ に対し，補助情報 $\mathcal{Y}_{t_n}$ が事前に割り当てられているものとする．このとき，学習データを用いて画像情報 $\boldsymbol{I}_i$ とその補助情報 $\mathcal{Y}_{t_i}$ の関係を学習することで，テスト画像 $\boldsymbol{I}_j$ に対してもその補助情報 $\mathcal{Y}_{t_j}$ を算出できる．テスト時には補助情報の集合の中から，テスト画像 $\boldsymbol{I}_j$ に最も近い補助情報を選択することで，その補助情報に関連付けられたラベル t_j を得ることができる．

3.2 提案手法の全体的な枠組み

提案手法では，画像特徴空間を補助情報と画像の比較空間とする．補助情報 $\mathcal{Y}_t$ にはクラス名 w_t とその定義文 $\mathcal{G}_t$ を用いる．提案手法では，まず意味特徴抽出器 φ によって補助情報を固定長のベクトル（以降，意味特徴と呼ぶ）に変換する．その後，次式に示す写像関数 F により意味特徴を画像特徴と同じ次元のベクトルに変換する．

$$F(\mathcal{Y}_t; \boldsymbol{W}_m, \boldsymbol{W}_\varphi) = \boldsymbol{W}_m^{(1)} f(\boldsymbol{W}_m^{(2)} \varphi(\mathcal{Y}_t; \boldsymbol{W}_\varphi)) \quad (1)$$

ただし，$\boldsymbol{W}_m$，$\boldsymbol{W}_\varphi$ は学習によって求まる重み行列であり，$f(\cdot)$ は非線形活性化関数の Rectified Linear Unit (ReLU) である．

重み行列は誤差逆伝播法によって更新され，その際に最小化する目的関数は次式である．

$$\frac{1}{N} \sum_{i=1}^{N} || \mathrm{CNN}(\boldsymbol{I}_i) - F(\mathcal{Y}_{t_i}; \boldsymbol{W}_m, \boldsymbol{W}_\varphi) ||^2 \quad (2)$$

ただし，$\mathrm{CNN}(\cdot)$ は事前に学習された畳み込みニューラルネットワークによって抽出された画像特徴であり，$F(\cdot)$ は式 1 に示した写像関数である．

テスト時には，次式に示すように比較空間上で最近傍探索を行い，テスト画像に尤もらしいクラスを割り当てる．

$$\arg\min_k || \mathrm{CNN}(\boldsymbol{I}_j) - F(\mathcal{Y}_{t_k}; \boldsymbol{W}_m, \boldsymbol{W}_\varphi) ||^2 \quad (3)$$

3.3 意味特徴抽出器

3.3.1 Simple Average モデル

Simple Average モデルでは，定義文中の各単語の埋め込みを平均することで得られるベクトルとクラス名の単語埋め込みを連結し，クラスを表現する意味特徴とする．

$$\varphi(\mathcal{Y}_t) = \left[v(w_t); \frac{1}{|\mathcal{G}_t|} \sum_{s=1}^{|\mathcal{G}_t|} v(w_s) \right] \quad (4)$$

ただし，$\mathcal{G}_t$ は定義文中に現れる単語の集合を表し，$v(\cdot)$ は事前に学習された Word2Vec によって得られる単語埋め込みを表す．なお，Simple Average モデルは学習によって更新されるパラメータはない．

3.3.2 Deep Average モデル

Deep Average モデルでは，定義文中の各単語の埋め込みを平均することで得られたベクトルに対し多層パーセプトロンによる特徴変換を施す．

$$\varphi(\mathcal{Y}_t; \boldsymbol{W}_\varphi) = \left[v(w_t); \boldsymbol{W}_\varphi^{(1)} f(\boldsymbol{W}_\varphi^{(2)} \frac{1}{|\mathcal{G}_t|} \sum_{s=1}^{|\mathcal{G}_t|} v(w_s)) \right] \tag{5}$$

3.3.3 Attention モデル

上述する 2 つのモデルは定義文中の各単語の重要度を考慮しないが，Attention モデルでは注意機構によって推定された単語重みによって定義文中の単語の重要度を陽に表現する．これにより，最終的な出力に貢献する単語の影響は強調し，貢献しない単語の影響は低減する効果を期待する．単語重みの推定には，定義文中のある単語の埋め込みとクラス名の単語埋め込みを入力とする多層パーセプトロンを用いる．

$$s_s = \boldsymbol{W}_\varphi^{(1)} f(\boldsymbol{W}_\varphi^{(2)} \left[v(w_t); v(w_s) \right]) \tag{6}$$

$$a_s = \frac{\exp(s_s)}{\sum_{k=1}^{|\mathcal{G}_t|} \exp(s_k)} \tag{7}$$

$$\varphi(\mathcal{Y}_t; \boldsymbol{W}_\varphi) = \left[v(w_t); \frac{1}{|\mathcal{G}_t|} \sum_{s=1}^{|\mathcal{G}_t|} a_s v(w_s) \right] \tag{8}$$

3.3.4 LSTM モデル

これまでに説明したモデルでは定義文中に含まれる単語の語順が特徴表現に与える影響は無い．ここでは，LSTM を用いることで単語の語順を反映した特徴抽出を試みる．具体的には，クラス名の単語埋め込みと LSTM による定義文の特徴表現を連結し，クラスの意味特徴とする．

$$\varphi(\mathcal{Y}_t) = \left[v(w_t); \mathrm{LSTM}(\mathcal{G}_t) \right] \tag{9}$$

ただし，LSTM(·) は LSTM の最後の隠れ状態とする．

4 実験

提案手法の有効性を検証するために行なった実験について説明する．実験において明らかにしたいことは次の二点である．

- クラス名の単語埋め込みに加えて辞書定義文から得た特徴表現を用いることは有効か
- 辞書定義文を用いる上で効果的な意味特徴抽出器はどれか

4.1 データセット

物体認識における標準的な大規模データセットである ImageNet を用いる．物体認識のコンペティションである ILSVRC 2012 で用いられた 1000 クラスを学習に用い，ILSVRC 2010 で用いられたクラスのうち ILSVRC 2012 に含まれるクラスを除いた 360 クラスについてテストした．

4.2 実験設定

4.2.1 意味特徴

ImageNet 上の物体クラスはそれぞれが WordNet 上の Synset（同義語グループ）に関連付けられている．ここでは Synset に含まれる全ての単語を Word2Vec によって分散表現とし，それらを平均したものクラス名を表す単語埋め込みとして用いる．Word2Vec には Wikipedia の全英文記事（2016 年時点）を学習した Skip-gram モデル [6] を用いた．クラス名を表す特徴表現と Synset に紐付けられた定義文を入力とし，3.3 章で説明した意味特徴抽出器によってクラスを表現する意味特徴を得る．4 種類の意味特徴抽出器を用いた提案手法に加え，クラス名の単語埋め込みのみを用いる Name only モデル，定義文中に含まれる各単語の埋め込みの平均のみを用いる Definition only モデルも合わせて実験した．

4.2.2 画像特徴

画像特徴として，CNN の一種である ResNet-152[5] の最後のプーリング層の中間表現を使用する．なお，この実験において，CNN のパラメータは ILSVRC 2012 の 1000 クラス物体認識のタスクを学習した際のパラメータであり，ファインチューニングは行わない．

4.2.3 評価指標

テスト画像 $\boldsymbol{I}_j$ が与えられた時，式 (3) によって算出されるスコアに基づいてテストラベルを順位付ける．この順位付け（ランキング）をモデルの予測結果とし

手法名	意味情報源 / CNN モデル	Hit@K	
		1	5
DEm [8]	N / G_B	11.0	25.7
Name only	N / R	11.5	25.7
Definition only	D / R	9.6	23.6
Simple Average	N + D / R	**13.0**	**28.8**
Deep Average	N + D / R	12.6	28.3
Attention	N + D / R	12.1	27.5
LSTM	N + D / R	11.6	26.2

表 1: ILSVRC 2010 \ 2012 における Hit@K
（N: クラス名，D: 定義文，G_B: GoogLeNet + バッチ正規化，R: ResNet）

て評価する．評価指標には Hit@K を用いた．これはモデルの予測結果のうち上位 K 件に正例クラスが含まれていれば正解とする指標であり，全テスト画像のうち正解した画像の割合を表す．

4.3 結果

表 1 に Hit@1 と Hit@5 の結果をまとめる．表 1 によると，**Name only** と **DEm** の性能は同程度である．Zhang ら [8] の大規模画像データセットを用いた実験においては，補助情報としてクラス名の単語埋め込みのみを想定している．この条件下において，DEm と Name only の 2 つの手法に本質的な違いため，妥当な結果である．また本実験は Zhang らの研究で実施された実験を正しく再現できていると考えられる．次に，単一の意味特徴を用いる場合において，クラス名と定義文を比較する．**Name only** と **Definition only** の結果より，クラス名を用いる方が定義文を用いるよりも良い性能を示すことがわかる．次に，クラス名と定義文を合わせて用いる 4 つの提案手法を比較する．クラス名と定義文を合わせて用いた手法はどちらかを単独で用いた手法よりも良い性能を示し，定義文を用いることの有効性が確認できた．4 つの提案手法の中では **Simple Average** が最も良い性能を示した．Simple Average は学習パラメータが少ないために過適合を抑えることができたと考えられる．

5 おわりに

本稿では，辞書定義文を用いてゼロショット物体認識を行う手法を新たに提案した．クラス名から取得した特徴表現と辞書定義文から抽出した特徴表現を合わせることで，画像との対応関係の学習に有効な特徴表現を得た．大規模なゼロショット物体認識の実験の結果，提案手法は既存手法を上回る性能を達成した．さらなる高精度化のためには，ゼロショット物体認識に適した正則化手法の導入や辞書定義文の構造を活用することが考えられる．

謝辞

本研究の一部は，文部科学省博士課程教育リーディングプログラム 早稲田大学「実体情報学博士プログラム」による補助，JSPS 科研費 (17H01831) の助成を受けた．

参考文献

[1] Z. Akata, S. Reed, D. Walter, Honglak Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of CVPR*, 2015.

[2] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of ICCV*, 2013.

[3] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proceedings of NIPS*, 2013.

[4] Tristan Hascoet, Yasuo Ariki, and Tetsuya Takiguchi. Semantic web and zero-shot learning of large scale visual classes. In *Proceedings of SNL*, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, 2016.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. 2013.

[7] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *Proceedings of ICLR*, 2014.

[8] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of CVPR*, 2017.

[9] X. Zhu, D. Anguelov, and D. Ramanan. Capturing long-tail distributions of object subcategories. In *Proceedings of CVPR*, 2014.