共起性を利用した物体認識における言語情報の有効性

黒澤 郁音　菊池 康太郎　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

本研究では物体認識のタスクを対象とし，画像中に映る物体の共起性を利用する枠組みにおいて，物体の名前から取得した言語情報の有効性を検証する．

画像認識の分野において，一枚の画像中に複数の物体が写っている場合における物体認識技術が多数提案されている [8, 5, 7]．これらは，それぞれの物体について，独立に識別を行うものがほとんどであった．近年ではこれに対し，画像中の物体間の共起性を考慮することで，物体の識別率を向上させる研究が盛んに行われている [10, 4]．共起性を考慮するために，複数の物体領域に対して物体クラスの同時確率を推定することを考えると，必要となるデータセットの量及び計算量が非常に大きくなる．そこで我々は前者の問題に対して言語特徴量を用いることで，多数の物体のラベルを少数の抽象概念で扱えるようにし，少量のデータセットでも共起性の学習を行えるようにした．また，後者の問題に対しては，条件付き確率場において用いられる平均場近似と [11]，近年提案されたガンベルソフトマックスによるサンプリング [2] を適用することで，これを解決している．本稿では，物体認識における言語特徴量の有用性を示し，平均場近似，ガンベルソフトマックスを適用したディープニューラルネットワークが正しく働くことを確認する．

2　問題とアプローチ

一枚の画像中の複数の物体を認識する手法 [8, 5, 7] の多くは，初めに複数の物体領域を検出し，各物体領域毎に領域内の画像特徴量を用いて物体クラスを推定するものであった．近年ではこういった手法を拡張し，物体間の共起性を利用することによって，より認識精度を高めようとする手法 [10, 4] が研究されている．物体の共起性とは，フォークと皿，キーボードとマウスのようないくつかの物体が画像中で共起しやすいという性質であり，それぞれ物体領域毎にクラス尤度を求めるより，全ての物体領域の物体クラスの同時確率を求めて物体認識を行った方が認識精度が高くなると考えられる．

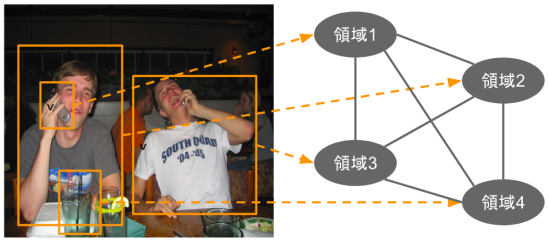

図 1: 物体領域候補をノードとする因子グラフ

そこで，事前に検出された物体領域をノードとするような，全てのノード間が結びついているペアワイズの因子グラフを考える．図 1 に物体領域候補をノードとする因子グラフを示す．この因子グラフ上の同時確率は，物体の領域の集合 A と因子関数 ψ によって以下のように表される．

$$p(x_1, x_2, ..., x_{|A|}) = \frac{1}{Z} \prod_{\{i,j|i,j\in A, i\neq j\}} \psi(x_i, x_j)$$
$$= \frac{1}{Z} \prod_{i,j} exp(f(x_i) + g(x_i, x_j) + f(x_j)) \quad (1)$$

このとき，Z は規格化定数であり，関数 f は独立に物体を推定する関数，関数 g は物体の共起性を求める関数である．このようなモデルを学習するためには二つの大きな問題が存在する．

一つ目の問題は，様々な物体の組み合わせにおける共起性を学習するために，大量の画像データセットが求められることである．すなわち，フォークとステーキの共起性を学習していても，スプーンとライスという未学習の組が与えられると，これを推定することができなくなるという問題が起こる．これはフォークとステーキの共起性から，食器と食品というある程度抽象的な概念での共起性を学習できていれば起こらないはずの問題である．そこで，言語的に類似した物体の

ラベルに対して類似したベクトルを与える手法として，SkipGram [6] を用いることで，抽象的な概念のレベルにおいて共起性を学習することを試みる．

二つ目の問題は，各領域における周辺確率を計算するために非常に大きな計算量を要するということである．この問題に対しては，マルコフ確率場で利用される平均場近似と，ガンベルソフトマックスによるサンプリング法を用いることで解決できる．

平均場近似は，同時確率分布を以下のように近似する手法である．

$$p(x_1, x_2, ..., x_{|A|}) \simeq \prod_{i \in A} p(x_i) \tag{2}$$

これによって，周辺確率を計算する必要がなくなる．この近似を行うためには，カルバック・ライブラー・ダイバージェンスを用いてこれらの分布の差を最小化すればよい．式 1 を式 2 のように近似するためには，以下のような式を繰り返し計算し，$p(x)$ を更新すれば良い．

$$\begin{aligned} p^0(x_i = k) &\propto exp(f_k(x_i)) \\ p^t(x_i = k) &\propto exp((|A| - 1)f_k(x_i) \\ &\quad + \sum_{j \in A \setminus i} \sum_{l \in Y} p^{t-1}(x_j = l) g_{kl}(x_i, x_j)) \end{aligned} \tag{3}$$

この式において，Y は物体クラスの集合である．

平均場近似によって計算量は大きく抑えられるが，式 3 を見ると，関数 $\sum_{l \in Y} p(x_j) g_{kl}(x_i, x_j)$ の計算量が大きいことが分かる．そこで，l をサンプリングすることによって，式 3 を近似することを考える．

$$\begin{aligned} &p^t(x_i{=}k) \\ &\propto exp((|A| - 1)f_k(x_i) + \sum_{j \in A \setminus i} \sum_{l \in Y} p^{t-1}(x_j{=}l) g_{kl}(x_i, x_j)) \\ &= exp((|A| - 1)f_k(x_i) + \sum_{j \in A \setminus i} E_{s \sim p^{t-1}(x_j)}[g_{ks}(x_i, x_j)]) \\ &\simeq exp((|A| - 1)f_k(x_i) + \sum_{j \in A \setminus i} \frac{1}{S} \sum_{s} g_{ks}(x_i, x_j)) \end{aligned} \tag{4}$$

ここで，S はサンプリング回数である．ニューラルネットワークなどを用いたとき，こういったサンプリングは微分不可能であるために，勾配を伝搬させることができなくなってしまう．これを解決するために，ガンベルソフトマックス [2] を用いた再パラメータ化トリックによって，微分可能なサンプリングを行う．

3 提案手法

式 1 をニューラルネットワークによって構成する．

関数 f は，対応する物体領域内の画像特徴量を入力とする全結合ニューラルネットワークとする．本研究では，物体領域を含む一枚の画像全体を物体識別タスクにおいて事前に訓練された VGGnet [9] に入力し，その畳み込み層から得られた特徴量マップに対し ROI プーリング [1] を施して得られた領域内の画像特徴量を入力として扱う．図 2 に，画像特徴量を入力とする関数 f の構成を示す．

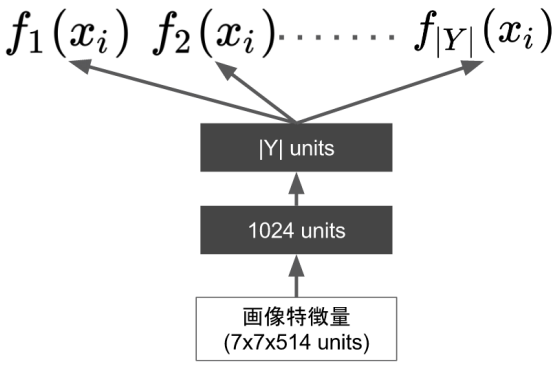

図 2: 画像特徴量を入力とする関数 f の構成

また，関数 g は，二つの言語特徴量と，ペアとなる二つの物体領域間の領域特徴量を入力とするニューラルネットワークとする．言語特徴量には，Wikipedia のコーパスを SkipGram によって学習した 300 次元のベクトルを用いる．また，領域特徴量は物体領域間の相対位置，面積比等を表した 4 次元のベクトルとする．図 4 に，領域特徴量の生成法を記述する．また，図 3 に，物体間の共起性を求める関数 g の構成を示す．

これらの関数を用いて，式 3 の通りに $p^t(x)$ を求める．すなわち，それぞれの物体領域に対し関数 f を用いてクラス尤度の初期値を推定し，関数 g を用いて互いの物体領域の共起性を元にクラス尤度を更新していく．更新を行った後のクラス尤度をモデルの出力結果とし，更新後のクラス尤度と正解として与えられるクラス尤度の交差エントロピーを損失関数とすることで学習を行う．したがって，関数 f と関数 g は同時に学習することができる．

4 実験

クラウドワーカーによって各物体に物体ラベルとその物体領域がアノテーションされた，VisualGenome [3] と呼ばれるデータセット中の画像 1,0805 枚を用いて実験を行った．そのうち，9,725 を学習用データ，1,080 を評価用データとして用いた．また，データセット中にアノテーションされた物体ラベルのうち，頻出順で上位 100 種の物体クラスを識別対象として用いた．

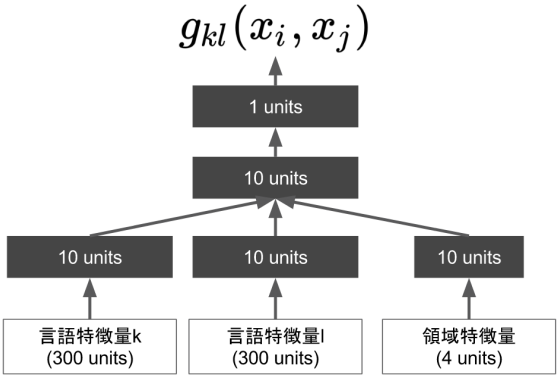

図 3: 言語特徴量と領域特徴量を入力とする関数 g の構成

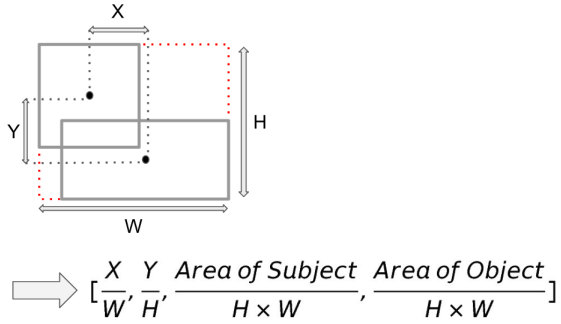

図 4: 領域特徴量の生成法

共起性の有効性を確認するため，各物体領域の物体認識を独立に行った場合と，平均場近似を用いた場合の認識精度を測定する．物体認識を独立に行うとは，式 3 において，p^0 を最終的なクラス尤度として扱うということである．また，言語特徴量の有効性を示すため，関数 g において one-hot なベクトルを入力とする場合，Skipgram のモデルから得られた言語特徴量を入力とする場合を比較する．

更に，物体領域をノードとするグラフに対し，画像特徴量のみを用いた手法 [10] もあわせて比較する．

5 結果と考察

表 1 に，平均場近似の有無，言語特徴量の有無による比較結果を示す．表より，平均場近似を用いることで物体の認識精度は向上していることが分かる．また，言語特徴量を用いることでさらに精度が向上することが分かった．更に，画像特徴量のみを用いた研究 [10] よりも性能が高いことが確認できた．

性能が向上した一方で，全体として物体認識精度が低いことが確認できる．原因の一つとしては，明確な区別の難しい物体クラスが多数存在することが挙げられる．表 2 に，区別の難しい物体クラスの例を示す．いくつかの物体は，`girl` と `woman` のような包含関係や，`leave` と `leaves` のような複数形の関係にあるため，クラスを統合することが求められる．

表 1: 提案手法における物体認識精度

手法	Precision
平均場近似無し	0.413
平均場近似有り (one-hot)	0.424
平均場近似有り (Skipgram)	**0.429**
Iterative Message Passing [10]	0.412

表 2: 明確な区別が難しい物体クラスの例

man	player
woman	girl
road	street
leave	leaves
window	windows
number	logo
leg	pants

6 まとめ

画像中に複数の物体が存在する場合において，物体同士の共起性を利用した物体認識を行うために，言語特徴量を用いる手法を提案した．

評価実験によって，各物体領域において独立に物体認識をするのに比べ，提案したモデルは高い識別性能を示すことが確認できた．また，言語特徴量を用いた場合に更に精度が向上することを確認した．

提案したモデルは，ガンベルソフトマックスによって計算量がある程度抑えられているが，各物体領域において独立に物体認識をするのに比べ，未だに計算量

が非常に大きい．今後は，この計算量を更に改善することで，大規模なデータセット，多くの物体クラスを用いて学習することを可能にする必要がある．

謝辞

本研究は JSPS 科研費 (17H01831) の助成を受けた．

参考文献

[1] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[2] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 32–73, 2017.

[4] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *arXiv preprint arXiv:1703.03054*, 2017.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*, 2015.

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

[10] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. *arXiv preprint arXiv:1701.02426*, 2017.

[11] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1529–1537, 2015.