言語特徴量を利用したシーングラフ生成の効率的な計算機構

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，与えられた画像に対するシーングラフを高精度かつ効率的に生成する手法を提案する．

シーングラフとは，図 1 に示すように，画像に含まれる物体とそれらの間の関係を記述するグラフである．画像の意味内容をコンパクトに表現できるため，画像検索や画像質問応答などの様々なタスクへの応用が期待されている．

我々は，高精度なシーングラフ生成を達成するために，画像中の物体のクラスを離散的な記号ではなく，類似度が定義可能な連続的な言語特徴量により表現することを提案する．これにより，学習データに存在しない未知の物体が認識時に現れた場合にも，類似する既知の物体と同様に扱うことが可能となる．

さらに，連続的な言語特徴量の活用を可能とするため，CRF とニューラルネットワークを組み合わせたアーキテクチャを導入する．このようなアーキテクチャでは計算量が問題となるが，多重線型な層を組み込むことにより，計算量を抑えつつ，最高水準の精度でシーングラフを生成することが可能となった．

2　シーングラフ生成における課題

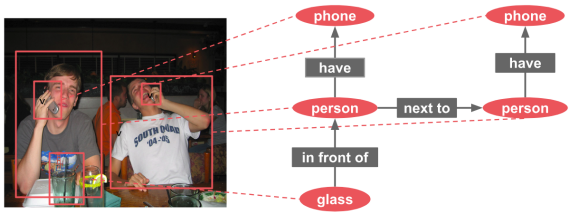


図 1: シーングラフの例

シーングラフは，図 1 に例を示すように，画像に映る物体をノードとし，それらの間に成立する関係を有向エッジとする有向グラフである．形式的には，⟨subject, predicate, object⟩ という形式を持つ三つ組を要素とする集合として定義することができる．ここで，subject は主体となる物体，object は客体となる物体を表し，predicate はこれらの物体間の関係を示す．図 1 の例には，⟨*person*, *have*, *phone*⟩ や ⟨*glass*, *in front of*, *person*⟩ などの三つ組が含まれる．

シーングラフ生成とは，与えられた画像に対して潜在的に存在するシーングラフを推定するタスクであり，画像中の物体の種別を認識し，かつ，それらの間の predicate を認識することが必要となる．

シーングラフ生成における大きな課題として，2 つの点が挙げられる．

一つは，物体認識と predicate 認識の双方を行うことが必要となることである．predicate 認識においてはそれに関わる物体のクラスが正しく認識されていることが望まれるが，物体認識を決定論的に行うことは難しく，物体クラスを定めるには逆に predicate の情報が有用である．この相互依存性を解決するためには，組み合わせ問題を効率的に解く必要がある．

もう一つは，学習データに存在しない物体が認識時に現れる可能性があることである．このような場合にも，それらの間の predicate を適切に認識する必要がある．従来のシーングラフ生成モデル [10, 9, 6, 2, 4] は，この課題に対する対処が不十分であり，predicate 認識のために極めて大きな学習データを必要とする．

本研究では，従来研究では十分に検討･達成されていない，これら二つの課題の同時解決について取り組む．

3　関連研究

従来のシーングラフ生成モデル [10, 9, 6, 2] はいずれも，画像特徴量を入力とした end-to-end なニューラルネットワークである． [10, 9, 2] らのモデルは，グラフ上のそれぞれのノードとエッジに対して特徴量を与え，隣接したノード，エッジ同士で互いに特徴量を伝搬し合う仕組みを持つ．彼らのモデルは我々のモデルと類似した構造を持つが，言語特徴量を導入する方法が自明でない．これに対し我々は，CRF の枠組みによって言語特徴量を容易に導入することを可能とする．

言語特徴量は三つ組を推定するタスクにおいて有効であることが，従来の研究 [4] より明らかとなっている．Lu ら [4] は CNN から得た画像特徴量のみを用いた場合と，言語特徴量 [5] を加えた場合で三つ組推定の精度を比較し，言語特徴量の有効性を明らかにした．これに対し我々は，シーングラフ生成でも言語特徴量を有効に活用させることが可能であるかどうかを明らかにする．

4 提案手法

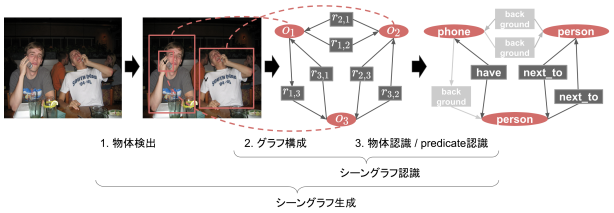


図 2: シーングラフ生成の流れ

図 2 にシーングラフ生成の流れを示す．

我々は従来の研究と同様にシーングラフ生成を二つの段階に分ける．

初めに，与えられた画像に対して N 個の物体領域候補を検出する．我々はこれらの物体領域候補に対応する確率変数を $O = o_1, o_2, ..., o_N$ とおく．このとき，それぞれの確率変数にどの物体クラスが当てはまるかは明らかではない．

N 個の物体領域候補が検出された後，各物体領域候補をノードとした完全グラフが構築される．このグラフにおけるエッジはそれぞれ物体間の `predicate` に対応する．我々はそれぞれのエッジに対応する確率変数を $R = r_{1,2}, r_{1,3}, ..., r_{N-1,N}$ とおいた．グラフ中の全ての確率変数 O, R に対して物体認識，`predicate` 認識を行うことで，シーングラフを完成させることができる．我々は，このタスクをシーングラフ認識と呼ぶ．また，この物体領域候補検出は既存の物体検出器 [1, 8] によって行うことができるため，シーングラフ認識のタスクにおいては物体領域候補検出の処理を除き，事前に正しい物体領域が検出されていることを前提とする．

同時確率 $p(O, R)$ は規格化定数 Z とエネルギー関数 $E(O, R)$ を用いて以下のように定式化される．

$$p(O, R) = \frac{\exp(-E(O, R))}{Z} \tag{1}$$

$$E(O, R) = -\sum_i \psi_u(o_i) - \sum_{i<j} (\psi_p(o_i, r_{i,j}, o_j) + \psi_p(o_j, r_{j,i}, o_i))$$

式中の ψ_u と ψ_p はエネルギー項である．ψ_u はグラフ中の各ノードに対応し，それぞれの物体領域候補について物体クラスを推定する．また，ψ_p はグラフ中の各三つ組に対応し，`subject`, `predicate`, `object` のクラスの組み合わせの妥当性を評価する．我々はこれらのエネルギー項を，図 3 と図 4 に示したニューラルネットワークによって構成する．

4.1 物体認識

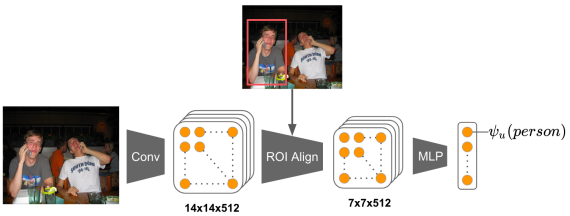


図 3: ノードに対応するエネルギー項 ψ_u

図 3 に示すように，エネルギー項 ψ_u は `Conv`, `RoI Align`, `MLP` の三つのモジュールによって構成されるニューラルネットワークである．

我々は `Conv` として，Faster-RCNN [8] において画像特徴量抽出の役割を担う畳み込み層を用いる．また，この畳み込み層はあらかじめ Visual Genome [3] で事前学習され，そのパラメータは更新されないよう固定される．我々は各物体領域候補毎の特徴量を得るために，`Conv` によって抽出された画像特徴量に対して `RoI Align` を用いる．

`MLP` は，`RoI Align` [1] によって得られた各物体領域候補毎の特徴量からそれぞれの物体クラスを認識するモジュールであり，ψ_u において唯一学習される部分である．その出力層は物体クラスと等しい数のユニットを持ち，それぞれのユニットはいずれかの物体クラスに対応する．

4.2 `predicate` 認識

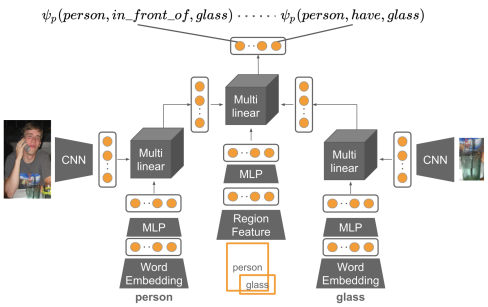
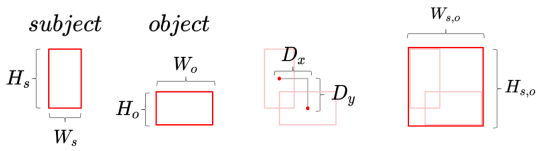


図 4: 三つ組に対応するエネルギー項 ψ_p

図 4 に示すように，エネルギー項 ψ_p は多重線型な層 (Multilinear layer) を含んだ MLP (multilayer perceptrons) によって構成される．多重線型な層とは，任意の数の実ベクトルを入力することができる，多重線型な関数である．例えば，三つのベクトル $x \in \mathcal{R}^{\mathcal{A}}, y \in \mathcal{R}^{\mathcal{B}}, z \in \mathcal{R}^{\mathcal{C}}$ を入力とする場合，多重線型な層の出力 $m \in \mathcal{R}^{\mathcal{D}}$ は重み $w \in \mathcal{R}^{\mathcal{A}\times\mathcal{B}\times\mathcal{C}\times\mathcal{D}}$ を用いて以下のように表される．

$$\begin{aligned} m_d &= Multilinear(x, y, z) \\ &= \sum_{a}^{A}\sum_{b}^{B}\sum_{c}^{C} w_{abcd} x_a y_b z_c \end{aligned} \tag{2}$$

このエネルギー項 ψ_p は subject，object それぞれの物体クラスを表す言語特徴量 [5, 7] と， それらの物体領域間の領域的関係を表す領域特徴量を入力とする．図 5 に示すように，領域特徴量は物体領域間の相対位置と被覆面積を 4 次元で表した実ベクトルである．



$$region\_feature = [\frac{D_x}{W_{s,o}}, \frac{D_y}{H_{s,o}}, \frac{H_s W_s}{area\_of\_union}, \frac{H_o W_o}{area\_of\_union}]$$

図 5: 領域特徴量の獲得方法

ψ_p の出力層は predicate と等しい数のユニットを持ち，それぞれのユニットはいずれかの predicate クラスに対応する．ただし我々は，何の predicate も当てはまらないことを表すための background という predicate を一つ加え，これに対応するユニットも同様に一つ加える．

4.3 シーングラフ認識

我々は二つのエネルギー項 ψ_u，ψ_p によって同時確率 $p(O, R)$ を計算するが，式 1 における規格化定数 Z の計算量は明らかに大きく，計算が困難である．

これに対して我々は平均場近似を適用する．平均場近似とは，近似的に各変数の周辺確率分布を求める手法であり，CRF の計算量を大幅に減らすことができる．物体の周辺確率 $p(o)$ と predicate の周辺確率 $p(r)$ は，以下のような計算を繰り返し行うことで計算される．

$$p^0(o_i) \propto \exp(\psi_u(o_i)) \tag{3}$$

$$p^{t+1}(o_i) \propto \exp(\psi_u(o_i) + \sum_{j \neq i} e_{obj}(i, j)) \tag{4}$$

$$p^{t+1}(r_{i,j}) \propto \exp(e_{pred}(i, j)) \tag{5}$$

$$\begin{aligned} e_{obj}(i, j) = &\sum_{o_j, r_{i,j}} p^t(o_j) p^t(r_{i,j}) \psi_p(o_i, r_{i,j}, o_j) \\ &+ \sum_{o_j \cdot r_{j,i}} p^t(o_j) p^t(r_{j,i}) \psi_p(o_j, r_{j,i}, o_i) \end{aligned} \tag{6}$$

$$e_{pred}(i, j) = \sum_{o_i, o_j} p^t(o_i) p^t(o_j) \psi_p(o_i, r_{i,j}, o_j) \tag{7}$$

これらの計算を T 回行った後得られた周辺確率 $p^T(o_i)$, $p^T(r_{i,j})$ を用いて，以下のように損失関数を定義できる．

$$L(O, R) = -\sum_{o \in O} \log\ p^T(o) - \lambda \sum_{r \in R} \log\ p^T(r) \tag{8}$$

この損失関数は物体クラスと predicate クラスについての交差エントロピー関数であり，ハイパーパラメータ λ は物体認識と predicate 認識の重みを調整する値となる．我々は，$T = 2$，$\lambda = 1.0$ とした．

以上のように平均場近似は計算量を大幅に減らすが，式 6 と式 7 において ψ_p の計算回数が非常に多く，依然として計算量が大きいことが分かる．ここで，ψ_u の多重線型性を有効に活用することで計算回数を減らすことができる．例えば，式 7 は以下のように計算できる．このとき，$f(x)$ は物体クラス x についての言語特徴量，$g(i, j)$ は物体ペア (i, j) に対する領域特徴量，h_{sbj} と h_{obj} はそれぞれ subject と object に対する画像特徴量である．

$$\begin{aligned} &\sum_{o_i, o_j} p^t(o_i) p^t(o_j) \psi_p(o_i, r_{i,j}, o_j) \\ &= \sum_{o_i, o_j} p^t(o_i) p^t(o_j) Multilinear(F(o_i), F(o_j), G(i, j), H_{sbj}, H_{obj}) \\ &= Multilinear(\sum_{o_i} p^t(o_i) F(o_i), \sum_{o_j} p^t(o_j) F(o_j), G(i, j), H_{sbj}, H_{obj}) \end{aligned}$$

$$F(o_i) = MLP(f(o_i)), \quad F(o_j) = MLP(f(o_j))$$
$$G(i, j) = MLP(g(i, j))$$
$$H_{sbj} = MLP(h_{sbj}), \quad H_{obj} = MLP(h_{obj})$$

5 実験

5.1 実験設定

我々は Visual Genome [3] と呼ばれる画像データセットを用いて評価実験を行った．このデータセットは 108,077 枚の画像を保持しており，各画像にはシーングラフがアノテーションされている．すなわちそれぞれの画像に対して，画像中の全ての物体について物体名称と物体領域がアノテーションされており，物体間には predicate がアノテーションされている．

Visual Genome の保持する画像のうち，75,631 枚の画像を学習用データセットとし，残りの 32,319 枚を評価用データセットとした．またこれらの画像にアノテーションされている物体名称と predicate のうち，150 の物体クラスと 50 の predicate クラスのみを扱う．この実験設定は [10] と全く同様である．

シーングラフ認識の精度の評価のために，Recall@K [4] と呼ばれる評価手法を用いる．これは，推定されたシーングラフを三つ組の集合に分解し，それらのうち上位 50 件の三つ組に対して Recall を測るものである．また，三つ組に対応するエネルギー項によって物体認識精度が向上する可能性があるため，物体認識の精度についても評価を行う．

表 1: 物体認識とシーングラフ認識における比較実験

	物体認識	シーングラフ認識	
モデル	識別率	R@50	R@100
CNN only	64.7	-	-
MESSAGE PASSING [9]	-	34.6	35.4
MOTIFNET [10]	-	35.8	36.5
Ours (one-hot)	64.9	35.6	36.3
Ours (Skip-gram)	**66.1**	**36.2**	**37.0**
Ours (poincare)	65.2	35.1	36.3

5.2 比較モデル

我々は，我々の提案モデルを含め，六つのモデルを比較した．まず，我々の提案モデルとして Ours (Skip-gram)，Ours (poincare)，Ours(one-hot) のシーングラフ認識精度を明らかにする．Ours (Skip-gram)，Ours (poincare) はそれぞれ，Skip-gram によって Wikipedia から得られた単語分散表現と，Wordnet から得られた poincare embedding [7] を言語特徴量とする．また，Ours(one-hot) は言語特徴量の代わりに物体クラスの one-hot 表現を扱う．これらのモデルを，従来の二つのモデル (MESSAGE PASSING，MOTIFNET) と比較する．また，物体認識のタスクにおいて，CNN のみを用いたモデル (CNN only) との比較を行う．

5.3 結果と考察

表 1 に評価実験の結果を示す．この表より，物体認識において Ours (Skip-gram) と Ours (poincare)，Ours (one-hot) が CNN only を越える精度を達成していることから，我々の提案したモデルが物体認識精度の向上に有効であることがわかる．また，物体認識とシーングラフ認識のいずれにおいても，Ours (Skip-gram) は Ours (one-hot)，Ours (poincare) を越える精度を達成しており，Skip-gram による言語特徴量が有効であることがわかる．更に Ours (Skip-gram) は，従来の研究で最高の精度を達成している MOTIFNET を含め，全ての既存のモデルより良いシーングラフ認識精度を得ており，我々の提案したモデルはシーングラフ認識においてな妥当なモデルであると考えられる．

一方で，Ours (poincare) は物体認識，シーングラフ認識の両方の精度において Ours (one-hot) を下回る．これは，poincare embedding を用いると共通の上位語を持つ単語の分散表現が極端に類似してしまい，モデルにとって区別できない物体クラスができてしまっているためと考える．

6 まとめ

本稿ではシーングラフ認識に対し言語特徴量を有効に活用するために，CRF とニューラルネットワークを統合したモデルを提案した．さらに，多重線型な層を導入することによって計算量の問題を克服した．実験では，我々のモデルが従来提案されてきたモデルを上回る精度を達成した．また，Skip-gram による言語特徴量が物体認識，シーングラフ認識において有効であることが明らかとなった．

今後は，Skip-gram 以外の手法による言語特徴量について検討するとともに，テキストコーパスから得られる物体名称の共起の統計の活用について研究を進める．また，どのような物体， predicate について単語分散表現が有効に働くのか分析を行なっていく．

謝辞

本研究は JSPS 科研費 (17H01831) の助成を受けた．

参考文献

[1] Kaiming He, et al. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988. IEEE, 2017.

[2] Matthew Klawonn and Eric Heim. Generating triples with adversarial networks for scene graph construction. *arXiv preprint arXiv:1802.02598*, 2018.

[3] Ranjay Krishna, 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] Cewu Lu, et al. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pp. 852–869. Springer, 2016.

[5] Tomas Mikolov, et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, pp. 2171–2180, 2017.

[7] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pp. 6338–6347, 2017.

[8] Shaoqing Ren, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

[9] Danfei Xu, et al. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2017.

[10] Rowan Zellers, et al. Neural motifs: Scene graph parsing with global context. *arXiv preprint arXiv:1711.06640*, 2017.