The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

物体ペアに対する画像領域群の関連性を用いた物体関係検出

Visual Relationship Detection Guided by Region-object Relevance

郷津　優介 *1
Yusuke Goutsu

*1産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

Visual relationships represent the semantic connection between observed objects in an image. Most of the existing works treat this task as a visual classification task. However, each type of relationships has a variety of object combination and each pair of objects has diverse interactions. In this work, we propose a region-object relevance guided framework to tackle this problem. More specifically, we consider only the relationships that can be occurred from image regions associated with the object pairs. The proposed method achieves the state-of-the-art on the Visual Genome dataset.

1.　はじめに

一般物体認識の分野では，Convolutional neural networks（CNNs）による特徴量抽出及び学習の発展 [Huang 17] により，静止画像に対するラベル推定のタスクから画像中の物体領域ごとにラベル推定する物体検出 [Ren 17]，画素ごとにラベル推定する物体領域抽出 [He 17] と呼ばれる高度なタスクへと発展してきている．この中でも物体検出は画像中に複数ある物体の位置とラベルを推定するもので，近年急速に発展してきている．この技術に加えて，並列処理で大量の演算を高速に行える GPU や大規模且つ良質なデータセットの貢献により多種類の物体でも高速に位置とラベルを特定できるほどになってきている．物体領域抽出以外の物体検出から発展した最近の研究として，Visual relationship detection (VRD) と呼ばれるタスクがある．Visual relationship とは，subject (sbj), predicate ($pred$), object (obj) のトリプレット $< sbj\text{-}pred\text{-}obj >$ で表現されるものである．VRD では物体検出による物体ラベル sbj, obj の推定に加えて，検出した物体ペアの関係ラベル $pred$ まで推定する．図 1 に VRD の具体例を示す．画像内の矩形で囲まれた複数の物体に対して，それらのペアの関係を表したトリプレットの候補が尤度の高い順に並べられている．

この物体関係を表すトリプレットの推定自体は新しいコンセプトの提案という訳ではなく，既に物体認識に言語処理を組み合せた先駆けの研究として Visual phrase detection (VPD) [Sadeghi 11] と呼ばれる研究がある．この中で複数の物体を含む矩形領域内での物体関係を記述した Visual phrase を導入しており，これは上記のトリプレットと同じものである．しかし，VPD では Phrase ごとにクラスを割り当てる必要があるために，物体と関係の種類が多くなると扱いにくくなるという欠点があった．一方，VRD では物体と関係のクラス分類問題にそれぞれ帰着できるために，[Lu 16] は扱える物体と関係の種類を拡張したデータセットを作成し，物体の位置とラベルの推定に加えて物体ペアの関係ラベルまで同時に推定している．

本研究では，この VRD というタスクを物体ペア $< sbj, obj >$ に対する関係 $pred$ に適切なラベルを割り当てる問題として目的関数の設計を行っている．また，設計における指針として，物体検出により得られる物体ラベルと画像内の物体領域の関連性に着目した．提案手法に対して，VRD 用のアノテーションを含む大規模データセットを用いた評価実験を行い，物体ペアに関連する画像領域から取り得る関係のみを考慮することが高い推定精度に繋がることを確認する．

連絡先: 郷津優介，産業技術総合研究所，〒 305-8560 茨城県つくば市梅園 1-1-1，Tel: 029-862-6345，Fax: 029-862-6552，Email: yusuke.goutsu@aist.go.jp

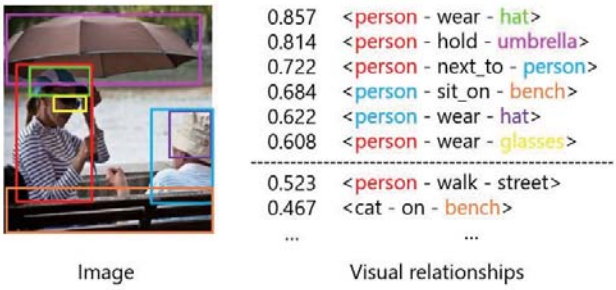

図 1: 物体関係検出

2.　関連研究

2.1　Phrase を利用した研究

トリプレットの推定自体は以前から他分野でも利用されている．[Sadeghi 11] 以外の Phrase を用いた研究として，物体間の意味的・幾何学的な関係を表現した 3D Geometric Phrase を導入することで，シーン認識，レイアウト推定，物体検出を同時に行った研究がある [Choi 13]．また，Relational Phraselets を用いて姿勢推定，フレーズ検出，Poselets のそれぞれの利点を組み合わせることで，行動認識，人物と物体の姿勢推定（オクルージョンにも対応）を行った研究がある [Desai 12]．

2.2　画像と言語の関連性を利用した研究

画像と言語の関連性は，Visual Question Answering (VQA) と呼ばれるタスクでも利用されている．このタスクでは，画像中の内容に関して質問文を与えると，その内容に関する答えを出力するものである．[Xu 16] では，質問文に対して画像中のどこに着目すれば答えを得られるかを示す Attention layer を導入したネットワークを用いている．Attention layer をスタックすることでより正確な着目位置を推定している．

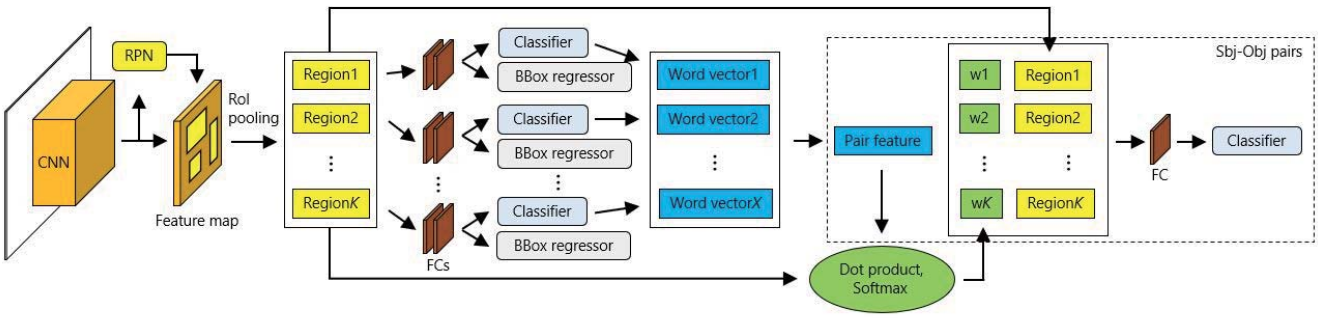

図 2: 提案システムの概観

3. 提案手法

本研究では，画像中にある物体の位置とラベルの推定に加えて物体ペアの関係ラベルまで推定する手法を提案する．図 2 に提案システムにおける全体の流れを示す．提案システムは，既存の物体検出用の Object detection (OD) モジュールに物体関係を推定する VRD モジュールを組み合せた拡張型となっている．ここで，既存の OD モジュールとして Faster R-CNN [Ren 17] を用いている．また，VRD モジュールでは，画像特徴と言語特徴の連想関係を構築し，画像中の物体ペアに関連する物体領域から起こり得る関係のみを推定している．

目的関数の最適化に関して，画像内で検出された物体ペアを $\boldsymbol{so} = \{s_m o_m; m = 1, ..., N\}$，それらの関係を $\boldsymbol{p} = \{p_m; m = 1, ..., N\}$ とすると，同時確率 $P(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o})$ を最大にする物体関係は以下の式で表される．

$$< \boldsymbol{s}^*, \boldsymbol{p}^*, \boldsymbol{o}^* >= \arg\max_{\boldsymbol{s},\boldsymbol{p},\boldsymbol{o}} P(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o}) \tag{1}$$

ここで，同時確率 $P(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o})$ を上記の OD モジュールと VRD モジュールによる推定結果に分割すると以下のようになる．

$$P(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o}) = P(\boldsymbol{so})P(\boldsymbol{p}|\boldsymbol{so}) \tag{2}$$

ここで，$P(\boldsymbol{so})$ は OD モジュールによる推定結果，$P(\boldsymbol{p}|\boldsymbol{so})$ は VRD モジュールによる推定結果になる．以下では各モジュールで算出される推定結果に関して詳細に説明していく．

3.1 物体ペアの検出

OD モジュールでは，物体ペアにおける各物体のラベル推定と画像中の位置同定を行う．物体ペアにおける各物体 sbj, obj の出現する確率が独立であると仮定すると，以下の式が成立する．

$$P(\boldsymbol{so}) = P(\boldsymbol{s})P(\boldsymbol{o}) \tag{3}$$

ここで，$P(\boldsymbol{s})$, $P(\boldsymbol{o})$ は OD モジュールにより得られる sbj, obj に属する物体がそれぞれ出力される確率である．

3.2 物体関係ラベルの推定

VRD モジュールでは，OD モジュールから出力される物体ペアのラベルを付与した時に，関連度の高い物体領域群から起こり得る関係ラベルを推定する．図 2 に示すように，OD モジュールから出力された物体ラベル群から $< sbj, obj >$ の物体ペアを選択し，ペア内での各ラベルの言語特徴を結合する．この操作は以下の式で表される．

$$\boldsymbol{l} = \mathrm{concat}(\mathrm{w2v}(s_m), \mathrm{w2v}(o_m)) \tag{4}$$

ここで，concat はベクトル結合の操作，w2v(·) は対象ラベルを分散表現したベクトルに変換することを意味する．この分散表現は，文章中である単語からその前後に現れる単語を予測できるような表現を学習することで得られるものである．

OD モジュールから得られる一定サイズに変換した物体領域候補群の画像特徴 $\boldsymbol{V} = \{\boldsymbol{v}_j | \boldsymbol{v}_j \in \mathbb{R}^M; j = 1, ..., K\}$（$K$ は物体領域候補の数，M は画像特徴の次元数）と上記の固定長の単語ベクトル $\boldsymbol{l} \in \mathbb{R}^L$ を用いて画像と言語の連想関係を学習することで，物体ペアに関連する画像領域を抽出する．即ち，物体ペアに対する j 番目の物体領域候補の重みは以下の式で計算される．

$$g_j = (\boldsymbol{A}\boldsymbol{v}_j + \boldsymbol{b}_A)^T(\boldsymbol{B}\boldsymbol{l} + \boldsymbol{b}_B) \tag{5}$$

$$w_j = \frac{\exp(g_j)}{\sum_k \exp(g_k)}, \boldsymbol{W} = \{w_j | j = 1, ..., K\} \tag{6}$$

ここで，(5) 式の $\boldsymbol{A} \in \mathbb{R}^{S \times M}$, $\boldsymbol{B} \in \mathbb{R}^{S \times L}$ は内積空間へ埋め込むための画像特徴，言語特徴に対する射影行列，$\boldsymbol{b}_A$, $\boldsymbol{b}_B \in \mathbb{R}^S$ はアフィン射影のバイアス項を表す（S は内積空間の次元数）．この内積は，物体領域候補と物体ペアの関係がベクトルの類似性に基づいて計算されることを意味する．(6) 式では，ソフトマックス関数により重み g_j の正則化（$\sum_j w_j = 1$, $w_j > 0$）を行うことで K 次元の $\boldsymbol{W}$ が取得される．これは物体ペアに対する物体領域候補群の関連度を表す．この関連度 $\boldsymbol{W}$ により物体ペアに対する関連領域マップを表すベクトルは以下の式で表される．

$$\boldsymbol{V}_W = \boldsymbol{W} \cdot (\boldsymbol{V} + \boldsymbol{b}_E) \tag{7}$$

ここで，$\boldsymbol{b}_E \in \mathbb{R}^{K \times M}$ はバイアス項を表す．これは画像と言語の連想関係により重み付けされた物体ペアに対する画像中の関連領域を表している．最終的に，物体ペア $s_m o_m$ が検出された時に起こり得る関係 p_m の条件付き確率は以下の式で計算される．

$$P(p_m | s_m o_m) = \mathrm{softmax}(\boldsymbol{W}_P \cdot f(\boldsymbol{V}_W) + \boldsymbol{b}_P) \tag{8}$$

ここで，f は ReLU と呼ばれる活性化関数，$\boldsymbol{b}_P \in \mathbb{R}^{C_p}$ はバイアス項を表す．また，$\boldsymbol{W}_P \in \mathbb{R}^{C_p \times M}$ により関連領域マップを表すベクトルを関係ラベルの種類の数 C_p と同じ次元数になるように変換している．(8) 式の計算を物体ペア数分だけ行うことで，(2) 式の $P(\boldsymbol{p}|\boldsymbol{so})$ が得られることになる．

3.3 マルチタスク損失の最適化

パラメータの学習は Faster R-CNN と同様にマルチタスク損失の最適化で行う．ただし，物体ラベルの推定と位置同定に

加えて (8) 式で計算される関係ラベルまで同時に推定する必要があるため，損失関数は以下の 3 項で定義される．

$$L(\boldsymbol{p}_{so}, u, \quad \boldsymbol{b}, \boldsymbol{t}^u, \boldsymbol{p}_p, v) = L_{sbj\text{-}obj}(\boldsymbol{p}_{so}, u) + I[u \geq 1] L_{bbox}(\boldsymbol{b}, \boldsymbol{t}^u) + L_{pred}(\boldsymbol{p}_p, v) \quad (9)$$

ここで，u, v は物体と関係の正解ラベル，$\boldsymbol{t}^u$ は正解バウンディングボックスの相対的な位置と大きさを表す．また，$I[u \geq 1]$ はアイバーソンの記法であり，$u \geq 1$ であれば 1，そうでなければ 0 となる．第 1 項と第 2 項の損失 $L_{sbj\text{-}obj}$, L_{bbox} は，元々の Faster R-CNN と同様に物体のクラス認識の損失，バウンディングボックスの回帰の損失を表す [Ren 17]．新しく加わった第 3 項の損失 L_{pred} は物体関係のクラス認識の損失であり，交差エントロピー誤差と呼ばれる真のクラス v に対する事後確率 p_p^v の負の対数として以下の式で表される．

$$L_{pred}(\boldsymbol{p}_p, v) = -\log p_p^v \quad (10)$$

(9) 式のマルチタスク損失を最小化するパラメータは，確率的勾配降下法により計算される．

4. 評価実験

4.1 実験設定

4.1.1 データセット

実験では，物体ペアとそれらの関係を推定するシステムの性能を評価するために Visual Genome Version 1.2 (VG) [Krishna 17] と呼ばれるデータセットを用いた．VG には 108K 枚の画像に 998K 個の物体関係を示すトリプレットが付与されており，トリプレットは 74,361 種類となる．クラウドソーシングを利用したアノテーション作業に伴う表記ゆれなどにより VG は整形して使用する必要があり，[Li 17] と同様に出現頻度の高い上位 150 種類の物体 ($C_{so} = 150$) と上位 50 種類の関係 ($C_p = 50$) を扱うことにした．ただし，物体ラベルには物体検出用に “背景” ラベルが，関係ラベルには物体関係検出用に “無関係” ラベルが追加で付与されている．前処理後には 95,952 枚の画像が残り，その中で約 5,000 枚をテストデータとして利用し，残りは訓練データとして利用した．

4.1.2 評価方法

実験では，提案システムの評価を行うために [Lu 16] と同様に以下の 3 つのタスクで先行研究との比較を行った．

Predicate Detection: 画像内にある物体の位置とラベル (sbj, obj) が与えられた状態で，物体ペアの関係ラベル $pred$ を推定する．

Phrase Detection: 画像が与えられた状態で，物体ペアを含むバウンディングボックスと物体関係を表すトリプレットのラベルを推定する．$<$ *sbj-pred-obj* $>$ 内の全ラベルが正解と一致し，推定された物体ペア (sbj, obj) を含むバウンディングボックスと正解のものの重なりが IoU > 0.5 であった場合に推定が正しいとする．

Relationship Detection: 画像が与えられた状態で，物体ペアのそれぞれの位置を示すバウンディングボックスと物体関係を表すトリプレットのラベルを推定する．$<$ *sbj-pred-obj* $>$ 内の全ラベルが正解と一致し，推定された物体ペア (sbj, obj) のそれぞれ位置を示すバウンディングボックスと正解のものの重なりが IoU > 0.5 であった場合に推定が正しいとする．

4.1.3 評価指標

アノテーション作業に伴う正解データの不完全性により，一般的に用いられる平均適合率 (mean average precision; mAP) は適切な評価指標にならない．実験では [Lu 16] と同様に評価指標として recall@50 (R@50) と recall@100 (R@100) を用いた．ここで，R@X はトリプレット推定結果の上位 X 個中に上記の評価指標による正解が含まれている数で計算される．

4.1.4 実行時の詳細

実験では，Caffe を用いて提案システムのネットワークを学習した．ネットワークは ImageNet で事前学習済みの VGG-16 モデルで初期化し，確率的勾配降下法を用いてネットワーク全体のパラメータを最適化している．ここで，学習率は 0.001 から段階的に減少させている．また，単語ベクトルの計算には Gensim 内にある単語ベクトル化手法を利用し，物体ラベルを 300 次元の低次元ベクトルで表現している（$L = 600$）．図 1 や 4.1.3 節からも分かるように，なるべくトリプレット候補の上位に正解が現れる必要があるため，物体ペアの関係中の “無関係” ラベルのスコアを用いることでヒューリスティックな候補削除を行っている．また，実行環境として GeForce GTX 1070 を用いた．

4.2 実験結果

提案手法を先行研究と比較した結果を表 1 に示す．ただし，他手法では 70,998 枚を訓練データ，25,000 枚をテストデータとして利用している．表より関係ラベルの推定では他手法に比べて最も高い性能を示していることが分かる．これは VRD モジュールにおいて画像と言語の連想関係により重み付けされた関連領域を利用したことで，正確な関係ラベルの推定に繋がったと考えられる．LP [Lu 16] でも単語ベクトルを用いているが，提案している線形写像モデルでは，多様な関係を推定するのに適切ではなかったと考えられる．また，ISGG [Xu 17] と比較した場合に，フレーズ検出では劣っているが，関係検出の R@50 に対しては高い性能を示している．提案手法がフレーズ検出や関係検出において比較的に性能が低いのは，他手法よりも未検出物体が多く存在することや不要なトリプレット候補の削除が有効にできていないことが原因であると考えられる．特に MSDN [Li 17] では通常の Faster R-CNN を拡張し，物体領域とキャプション領域の候補を推定する Region proposal network (RPN) を組み合せることで，文脈まで手掛かりにした物体検出を行っている．提案手法の OD モジュールを同手法に置き換えることで，フレーズ検出や関係検出の精度がさらに向上できると期待される．

提案手法と MSDN [Li 17] におけるテスト時の処理速度を比較した結果を表 2 に示す．どちらも同じ環境で実行している．ここで，表の数値は画像入力から RPN による候補領域の取得までの時間は除外されている．ただし，提案手法における括弧内の数値は画像の入力からトリプレット候補が出力されるまでに掛かった時間を示している．表より提案手法の方が処理速度は 1.7 倍速いことが分かる．さらに，実際の画像入力からトリプレット候補出力までの実行時間は物体検出のみと比べて 200ms しか変わらないことが分かる．これらの結果より end-to-end に設計したネットワークアーキテクチャが有効であったと考えられる．また，MSDN [Li 17] では提案手法と同様に Faster R-CNN を用いているが，上記の議論により複数の RPN を取得する必要があるために，除外されている部分は通常よりもさらに時間が掛かると考えられる．

最後に，提案手法による物体関係検出の結果を図 3 に示す．検出された物体ペアごとに関係ラベル候補のトップ 5 に正解が含まれていたものは色付きで示している．図中の “clouds”, “cap”, “clock”, “car”, “window”, “ground” は検出されなかったことからも，物体検出の精度向上が必要であると分かる．

表 1: VG を用いた場合の各タスクでの先行研究との比較結果

Comparison	Predicate Detection		Phrase Detection		Relationship Detection	
	R@50	R@100	R@50	R@100	R@50	R@100
LP [Lu 16]	26.67	33.32	10.11	12.64	0.08	0.14
ISGG [Xu 17]	58.17	62.74	18.77	20.23	7.09	9.91
MSDN [Li 17]	67.03	71.01	**24.34**	**26.50**	**10.72**	**14.22**
Ours	**67.52**	**74.57**	11.62	12.87	7.48	8.27

表 2: テスト時の処理速度の比較結果

Time (ms)	Ours	MSDN [Li 17]
VRD	182 (378)	309
OD	- (171)	-

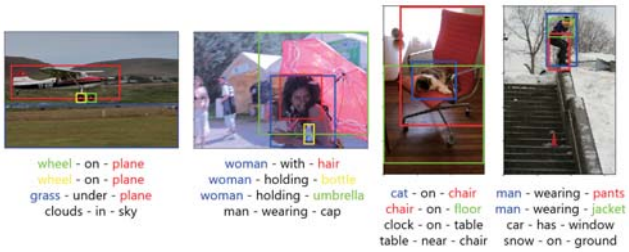

図 3: 物体関係検出の結果

5. 結論

本研究では，画像中に複数ある物体の位置とラベルの推定に加えて物体ペア群の関係ラベルまで推定する物体関係検出を扱った．従来の画像認識に沿った手法では，このタスクに対して物体と関係の種類が多くなると扱いにくくなるという欠点があった．しかし，物体検出が発展してきたおかげで，物体と関係のクラス分類問題にそれぞれ帰着できるようになった．本研究では，このタスクを物体ペアに対する関係に適切なラベルを割り当てる問題として目的関数の設計を行い，設計における指針として，物体検出により得られる物体ラベルと画像中の物体領域の関連性に着目した．提案手法に対して物体関係検出用のアノテーションを含む大規模データセットを用いて評価実験を行い，物体ペアに関連する画像領域から取り得る関係のみを考慮することで，一部のタスクおいて先行研究に比べて高い精度が実現できることや画像ごとの実行速度も 0.4 秒以下であることを確認できた．

本研究を進めるにあたり，全く新しい研究分野であったために，修士学生であった鈴木哲平氏には多くの知識や示唆を頂き，大変お世話になりました．ここに感謝の意を表します．

参考文献

[Huang 17] Huang, G. and Liu, Z. and Weinberger, K. Q. and van der Maaten, L.: Densely Connected Convolutional Networks, in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, (2017)

[Ren 17] Ren, S. and He, K. and Girshick, R. and Sun, J.: Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks, in *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* Vol. 39, No. 6, pp. 1137–1149, (2017)

[He 17] He, K. and Gkioxari, G. and Dollár, P. and Girshick, R.: Mask R-CNN, in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 2980–2988, (2017)

[Sadeghi 11] Sadeghi, M. A. and Farhadi, A.: Recognition using Visual Phrases, in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1745–1752, (2011)

[Lu 16] Lu, C. and Krishna, R. and Bernstein, M. and Fei-Fei, L.: Visual Relationship Detection with Language Priors, in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 852–869, (2016)

[Choi 13] Choi, W. and Chao, Y. W. and Pantofaru, C. and Savarese, S.: Understanding Indoor Scenes using 3D Geometric Phrases, in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33–40, (2013)

[Desai 12] Desai, C. and Ramanan, D.: Detecting Actions, Poses, and Objects with Relational Phraselets, in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 158–172, (2012)

[Xu 16] Xu, H. and Saenko, K.: Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering, in *European Conference on Computer Vision (ECCV)*, pp. 451–466, (2016)

[Krishna 17] Krishna, R. and Zhu, Y. and Groth, O. and others: Visual Genome: Connecting Language and Vision using Crowdsourced Dense Image Annotations, in *International Journal of Computer Vision (IJCV)*, Vol. 123, No. 1, pp. 32–73, (2017)

[Xu 17] Xu, D. and Zhu, Y. and Choy, C. B. and Fei-Fei, L.: Scene Graph Generation by Iterative Message Passing, in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5419, (2017)

[Li 17] Li, Y. and Ouyang, W. and Zhou, B. and Wang, K. and Wang, X.: Scene Graph Generation from Objects, Phrases and Region Captions, in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 1261–1270, (2017)