シーングラフ抽出に向けた再帰的ニューラルネットワークとトリプレットによるクラス数についてスケーラブルな物体間関係集合認識

中山研究室 修士二年 増井 建斗

1 概要

風景や物体などの一般的な環境を撮影した画像から，画像内の情報を物体間関係性の集合として出力する手法を提案する．本論文では画像内に含まれれる物体同士の関係を認識することを物体間関係認識と呼び，さらに物体間関係の集合を出力することを物体間関係集合認識と呼ぶ．タスクとしては一般画像からその画像内に含まれる物体と，それら物体同士の関係性の集合を出力する．

物体間関係性に関する問題を機械学習で取り扱う場合，そのクラス数が大きな問題となる．関係性を fact と呼び，fact を構成する要素を subject, predicate, object とする．(subject, predicate, object) のタプル (トリプレット) として考慮しなければならない fact のクラス数 N_f は各要素それぞれのクラス数 (N_s, N_p, N_o) の積となり，それぞれの増加と共に爆発的に増加してしまう．既存研究による物体間関係性認識では構造学習とランキング学習による手法が一般的で，それらは全ての fact に対する尤度を表すスコアを計算する．この時，N_f の爆発的な増加と共に計算量が大きく増加してしまう問題がある．さらに，スコアに基づいたモデルは画像に関する説明的な関係性の集合を出力するには適していないという問題がある．

本研究では再帰的ニューラルネットワーク (RNN) とトリプレットユニットを用いたモデルを提案し，クラス数に関する計算量の問題を解決する．さらに，RNN の学習に用いる損失関数を画像に対する関係性の集合全体に対して設計することで，ベースラインとしたモデルと比較してより様々な単語を用いた関係性の集合を出力する．

2 関連研究

物体間関係集合認識の関連研究として，ランキング学習による物体間関係認識 [4, 2, 1] がある．しかし，ランキング学習による既存手法では考えられる全ての fact のスコアを計算する必要があり，fact の数に関するスケーラビリティが課題となる．もう一つの問題として，シーングラフを構成するような物体間関係集合認識のためには出力される集合全体の性質を捉える必要が有るのに対して，既存手法は一つの関係性にのみ注目してスコアを計算している．

3 提案手法

クラス数に関するスケーラビリティを確保するために全ての fact に関するスコアの計算を避け，更に集合全体に対して最適化を行うために，提案手法として再帰的ニューラルネットワークとトリプレットユニットによるモデルを図 1 に示す．

提案手法は再帰的ニューラルネットワークの内部状態を更新しながら逐次的に fact を出力し，最終的な集合の出力とする．各出力時に内部状態を物体間関係性に変換する部分をトリプレットユニットと呼び，本論文では 2 つを提案する．ひとつは内部状態を一度に物体間関係性に変換する Subject-Predicate-Object at once(SPO)．もう一つは内部状態から 2 つの物体認識を行ったあと，物体認識の結果から関係性を予測するモデル Subject-Object then Predicate(SOP) である。

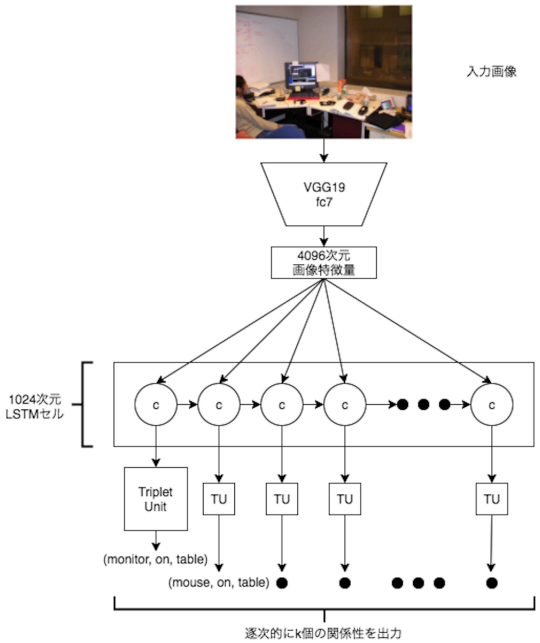

図 1: 再帰的ニューラルネットワークとトリプレットユニットによる提案手法

表 1: 提案手法と関連研究手法の，k 個の関係性出力に必要な計算量の比較．M は物体検出器の検出した領域数，N は $N \approx N_s \approx N_p \approx N_o$ として示している．

手法	計算量
MLP-Rank (ベースライン)	$O(kN)$
SPO (提案手法 1)	$O(kN)$
SOP (提案手法 2)	$O(kN^2)$
Visual Phrases[4]	$O(kMN^3)$
VRD[2]	$O(kM^2N^3)$
SSVM[1]	$O(kM^2N^2)$

最大 k 個の fact を集合として出力する際の，提案手法による計算量と既存手法の計算量を表 2 に示す． 関連研究は物体領域情報を用いているなど, 利用されている特徴量と計算量が提案手法と大きく異なるため, 性能をそのまま比較することが妥当ではない. 提案手法と同等の特徴量と計算量を持ったベースラインとなる手法として,MLP-Rank を実装した.MLP-Rank は関連研究で用いられたモデルから物体領域情報を取り除き, 計算量について同等となるように改変を行ったものである.

4 実験

実験の目的は主に 3 つである．一つはモデルのクラス数に関するスケーラビリティの確認．二つ目はベースラインと提案手法の出力する集合の性質の違いの確認．そしてトリプレットユニットの汎化

表 2: 提案手法と関連研究手法の，k 個の関係性出力に必要な計算量の比較．M は物体検出器の検出した領域数，N は $N \approx N_s \approx N_p \approx N_o$ として示している．

Model	Cost
MLP-Rank (Baseline)	$O(kN)$
SPO (Proposal 1)	$O(kN)$
SOP (Proposal 2)	$O(kN^2)$
Visual Phrases [Sadeghi 2011]	$O(kMN^3)$
VRD [Cewu 2016]	$O(kM^2N^3)$
SSVM [Atzmon 2016]	$O(kM^2N^2)$

表 3: CSK1000 に出現する一枚の画像に対する，出力された集合含まれる単語の種類数の平均．top-k の $k = 20$ としたときの出力結果．

Model	MLP-Rank	SPO	SOP	正解ラベル
平均単語種類数	11.0	14.4	15.9	15.6

表 4: $\overline{|facts|/I}$ は各画像にラベル付された fact の数の平均．N_f は考慮しなければならないクラス数．N'_f はデータセットに実際に含まれていた fact のクラス数．

name	$\overline{\|facts\|/I}$	N_f	N'_f	N_s	N_p	N_o
CSK5	1.21	**125**	50	5	5	5
CSK100	4.11	**850000**	30117	100	85	100
CSK1000	10.51	**833918730**	253764	894	985	947

性能の確認である．

スケーラビリティの確認のためにクラス数が可変なデータセット (CSK) を作製し，様々なクラス数での実験を行った．CSK の詳細を表 4 に示す．Precision@k と Recall@k の評価結果を図 2 に示す．この図から，k がデータセットの平均 fact 数に近い範囲では，クラス数が増えた場合に計算量を抑えた提案手法がベースラインよりも良い性能を出していることが確認できる．ランキング学習を用いたモデルは k を増やすほどに Recall@k が増加するが，提案手法はその傾向が抑えられている．これは，提案手法はデータセットで示された物体間関係集合と同様の数の関係性を出力するように最適化されており，それ以上の出力数については重複が多くなるためである．次に 2 つのトリプレットユニット,SPO と SOP を比較すると,SOP が最もよい汎化性能を出していることが確認できた．

図 3 と図 4 が，モデルに対する入力画像と出力された集合の例である．これらの結果から，提案手法とベースラインの出力の性質に違いがあることがわかる．教師ラベルは手作業でラベルを作製する際に，同じ単語をなるべく用いず，違った表現で物体間関係を追加するように作製されている．このため，ラベル集合は各グラフ同士の接続が少なく，多くの単語が用いられている．定量的な評価のため，一枚の画像の表現に用いられる単語の種類数の平均を表 3 に示す．これに対してランキング学習による出力では，物体間関係集合の性質を考慮しないために，比較的少ない種類の単語の組み合わせで集合を構築していることがわかる．提案手法ではベースラインと比較して，より多くの単語を用いている．

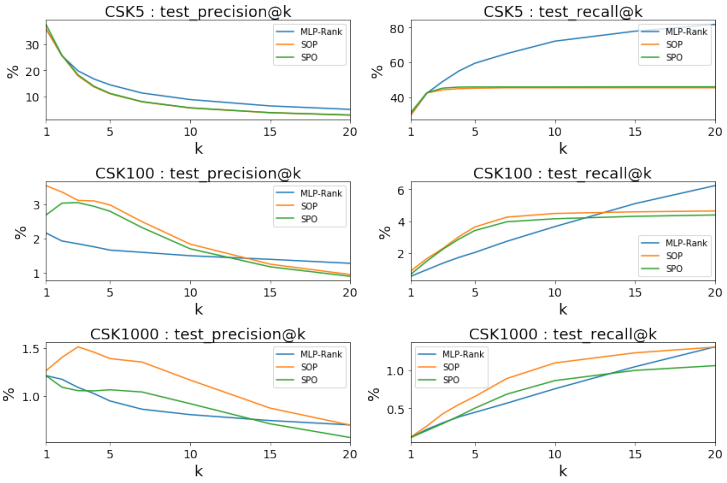

図 2: クラス数を増やしていった際の Precision@k と Recall@k

図 3: 入力画像

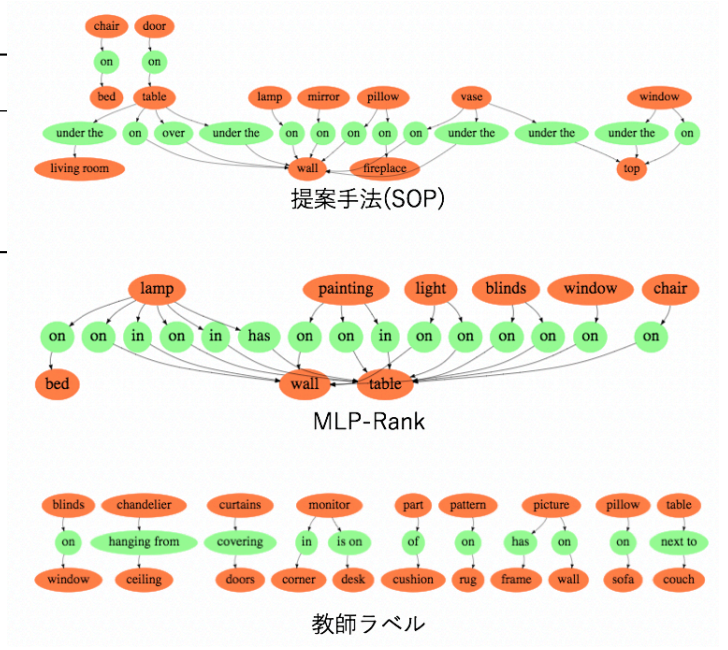

図 4: 出力された物体間関係集合と教師ラベル

5 結論

本論文では，一般画像を説明するデータ構造であるシーングラフ [3] を画像から検出するという最終目的に向けて，シーングラフの一部である物体間関係集合の認識というタスクに取り組んだ．物体間関係認識を目的とした既存研究の，全ての fact のランキングを軸にした手法に対して，シーングラフ検出を前提においた物体間関係集合認識を行う提案手法によって本論文は次の貢献を行った．

- 物体間関係集合出力を目的として RNN を応用しトリプレットユニットを用いたモデルを提案
- クラス数爆発に伴う計算量問題をトリプレットユニットを用いて解決
- 基本的なトリプレットユニットより汎化性能を持ったものとして SOP を提案
- 関係集合に対して最適化することで，より多くの単語を含むシーングラフに近い関係集合を出力

以上が，本論文の結論である．

参考文献

[1] Yuval Atzmon et al. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016.

[2] Lu Cewu et al. Visual Relationship Detection with Language Priors. *ECCV*, pages 852—-869, 2016.

[3] Justin Johnson et al. Image Retrieval using Scene Graphs. *CVPR*, pages 3668—-3678, 2015.

[4] Mohammad Amin Sadeghi et al. Recognition using visual phrases. In *CVPR*, pages 1745–1752. IEEE, 2011.