Visual Scene Graph Extraction

中山研究室 修士二年 増井 建斗

2017 年 1 月 4 日

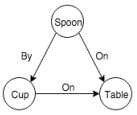

図 1　シーングラフの例

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，画像の内の情報を Scene Graph[2] と呼ばれるグラフ構造として出力するモデルを教師あり学習によって学習する方法を提案する．Scene Graph は画像内の情報を，物体とその属性のペアによるグラフ構造で表現したものである．教師あり学習に必要な画像と Scene Graph は制作に多くの手間がかかり，学習に十分なデータ・セットが用意されていないため，コンピュータグラフィックスによって自動生成したデータ・セットを用いて学習を行う．また，Scene Graph を出力可能な機械学習モデルの提案を行う．

2　背景

家庭内で家事を行うロボットの意思決定や画像検索においては，画像内の物体認識に加えて，それら物体同士の位置関係などの 2 物体以上に渡る関係性も重要な情報となる．従来の画像認識では画像内の物体のラベル付けと位置検出が行われているが，物体同士の関係性を含む情報を抽出する試みには，多くの研究余地が残されている．この研究の最終的な目標は，画像から物体同士の関係を含む情報を抽出することで家事ロボットによる意思決定を補助することであるが，画像検索などの問題にも適用可能である．

3　問題設定

Scene Graph では，画像内の情報を述語とその引数で表現する．例えば，Spoon，Cup，Table が存在するとき，Cup on Table という情報は On(Cup,Table) のように表現され，On が述語 (Predicate)，Cup,Table がその引数 (Parameter) である．述語と引数のペアを一つの事実 (Fact) と呼ぶ．この時，述語をエッジ，引数をノードと解釈すると画像内の情報が事実の集合，グラフ構造で表現され，全体としては図 1 のように表現される．この研究では，画像からこの Scene Graph を出力することを目標とする． この研究には 2 つの課題がある．一つは画像からのシーングラフの抽出．もう一つは学習に必要なデータセットの準備である．2 つの課題のうち，データセットについてはコンピューター・グラフィックス (CG) を用いてデータセットを作製することで解決することとした．残る問題については，事実の抽出が可能であるかの検証を行ったうえで，シーングラフの抽出を目指すこととした．以降，データセットの作成と，事実の抽出，シーングラフの抽出について説明する．

4　関連研究

関連研究として Visual Relationship Detection with Language Priors[4] があり，彼らは構造学習に近い手法で一般画像からの Scene Graph の抽出に取り組んでいた．構造学習に近い手法を用いて画像から Scene Graph を抽出する場合，可能性のある Scene Graph 全体の中から最適なものを選んで出力するために，出力候補すべてについてのスコアを計算する必要がある．Scene Graph を構成する Fact の数は, オブジェクトの種類数を M, 関係性の種類数を N とした時に $M \times N \times M$ となる．Scene Graph は任意数の Fact の集合であるため，スコアを計算する必要がある Scene Graph の数が膨大となってしまう問題がある．本研究では Recurrent Neural Network (RNN) を用いて Scene Graph を抽出することで出力候補全てについての候補を計算する必要性を取り除く．

5　データセットの作成

以降で提案する教師あり学習モデルのため，大量に画像と Scene Graph のペアを自動生成している．Graph 学習用のデータ・セットは実世界データ，CG データ共に現在製作中であり，10 月末に完成する見込みである．

6　事実の抽出

画像認識においては，畳込みニューラルネットワークが画像の特徴獲得に有効であることが確認されている．畳み込みニューラルネットワークは一般的には識別器と組み合わされ，画像の識別で高い精度を上げている．一般的な画像認識では一つの入力画像からその物体の種類を答えるが，今回は一つの事実を出力するために工夫が必要となる．画像特徴量の抽出には一般的な畳込みニューラルネットワークを利用しているが，出力が三叉に分岐している．それぞれが 2 つの物体とその関係に対応しており，この 3 つの出力を一つの事実として解釈する．この学習モデルを利用して事実の抽出を学習したところ，物体 (述語引数) に関しては 80 から 90% 以上の精度，物体間関係（述語）については 95% 以上の精度で正解することができた．

7　シーングラフの抽出

シーングラフの抽出では，複数の事実を出力させる必要がある．一つの入力から複数の時事を出力させるため，Long Short Term Memory (LSTM)[1] を利用する．LSTM はニューラルネットワークの一種だが内部に状態を持ち，一回の出力ごとに内部状態を更新し可変長の出力が行える．一般的には自然言語処理等，前の出力が次の出力に影響を与える問題において力を発揮するが，今回はこの可変長出力部分を利用するために用いている．また，Scene Graph においてはある Fact が他の Fact の存在に関与する場合も多くあるため，前

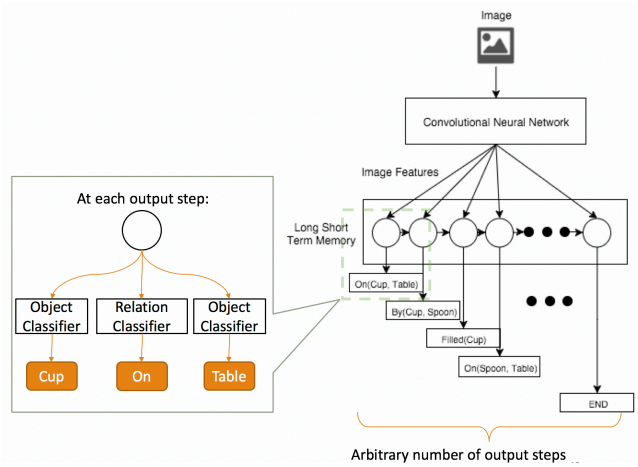

図 2　シーングラフ抽出モデル

図 3　学習中のモデルと損失関数の出力

述の性質は有用であると考えられる． 図 2 にモデルの全体構造を示す．画像の入力から Scene Graph の出力までの流れは，次のようになる．まず画像を幾層にも重ねられた畳み込みニューラルネットワークに入力し，その特徴表現を出力させる．次に，得られた特徴表現を LSTM に入力し，LSTM が終了記号を出力するまで，述語と引数のペアを出力させる．最終的に得られた述語と引数のペアの集合を，画像の Scene Graph とする．

8　実験

CG によるデータ生成が未だ終わっていないため，Visual Genome[3] という一般画像に Scene Graph が付属しているデータ・セットと提案したモデルを用いて実際に Scene Graph の出力を行った．モデルがデータ・セットに対して十分な表現力を持っていることを確認するため，学習に用いるデータ・セットの数を 100 枚と 5 万枚の 2 通りで実験を行った．小さなデータ・セットに対して過学習できなければ，より大きなデータ・セットに対して十分に学習できないと考えられるためである．この実験では最終層とその直前の LSTM のパラメータのみを学習さえ，畳込みニューラルネットワークのパラメータは一般画像認識タスクで事前に学習したもので固定している．

学習中のモデルの損失関数の移動平均を図 3 に示す．100 枚の画像で学習したモデルは損失が約 5，全画像 (5 万枚) で学習したモデルは約 1 5 程度で収束していることが確認できる．どちらの例も損失が大きく教師データに対する完全な学習は出来ていないことがわかる．

図 4 と表 1 に，テストデータセットの中からモデルに入力した画像と，その出力 Scene Graph，正解の Scene Graph を示す．提案したモデルは表の上から一つずつ Fact を出力している．100 枚の画像で学習したモデルは複数の種類の Fact を出力しているものの，”bus on bus”等の一般に起こり得ず，かつデータ・セットに含まれない Fact を出力してしまっている．反対に全画像を用いて学習したモデルは，”window on bus”など，学習データに多く含まれていた Fact

図 4　入力した画像

表 1　モデルの出力と正解の対応

100 Images	Whole Images	True Label
Bus on man	Window on Building	Sign on bus
Sign on building	Window on bus	Sign on front of bus
Bus on bus	Window on bus	Sign on front of bus
Sign on man	Window on bus	Bus stop on a sidewalk
Bus on building	Window on bus	Vehicle on a street
Bus on bus	Window on bus	End of a vehicle
Bus on man	Window on bus	Lamp post on a street
Sign on man	Window on bus	Bus going down street
Bus on building	Window on bus	Sidewalk of a city street
Bus on bus	Window on bus	Building at end of road
Bus on building	Window on bus	Vehicle behind bus

を偏って出力する傾向が確認できた．これらの結果は，Scene Graph の認識において重要な 2 つの要素が提案モデルでは未だ考慮されていないためだと考えられる．一つは物体の領域検出が行われていない点，もう一つは 3 つの出力層がお互いの出力について影響を受けない点である．前者は、モデルが各ステップにおいてどの事実を出力するかの決定に必要となる．後者は、”bus on bus”等のように 3 つの各出力層が最頻出なクラスを出力してしまうことを抑制し、3 出力間の共起を考慮した出力を行うために必要だと考えられる．

9　結論と今後の課題

家事ロボットの行動決定に必要な Scene Graph Extraction に向けて，教師あり学習の枠組みで画像から Scene Graph を出力するモデルを提案し，先行研究で必要となり問題だった解空間の全探索処理を排除している．提案したモデルには多くの改善の余地が残っており，引き続き実装を進める必要がある．Visual Genome による実験を行ったが，家庭内ロボットのための環境認識の学習に適したデータ・セットとして，CG によるデータ・セット作製を引き続き行っている．

参考文献

[1] Hochreiter et al. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[2] Johnson et al. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015.

[3] Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[4] Lu et al. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.