意味構造に着目したシーングラフ生成手法の提案

筑波大学
図書館情報メディア研究科
2020年3月
嵐　一樹

目 次

図 目 次

第1章　はじめに

画像内容を認識し画像情報を構造的に理解することは，コンピュータビジョンの分野において重要な課題となっている．画像情報を構造的に理解するというのは，画像中に示されている物体同士の関係や物体の状態などを，人間が視覚的に判断できるのと同等に理解するということである．本稿では画像を構造的に理解するため，画像キャプションとシーングラフを用いる．キャプションとシーングラフについては後述する．本章では，近年の画像認識に関する研究に触れつつ，研究の概要について述べる．

近年，計算資源の発達や深層学習の出現により，機械学習を用いた様々なタスクにおいて大きな進歩が見られる．画像認識の分野に絞ると，物体検出 [1, 2, 3, 4] やシーングラフ [5, 6, 7, 16, 17, 18]，画像キャプション生成 [11, 12, 14] などの研究がある．これらの研究はどれも画像の理解を目指す内容となっている．物体検出は，画像情報から特徴量を抽出し，画像中の物体カテゴリや物体位置などを特定する分野である．画像キャプション生成とシーングラフに関しては後述するが，概要を述べると，どちらも画像情報を元にして意味構造を自然言語で出力する．

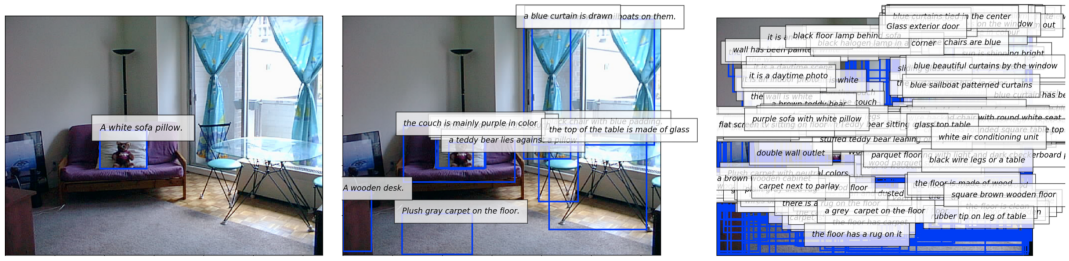
画像キャプションとは，自然言語での画像情報の短文説明のことである．画像キャプションを生成することを画像キャプション生成と言い，広く研究されている．画像キャプション生成では主に，画像を入力として自然言語のテキストを出力する．画像特徴量を抽出し，それらを入力としたニューラルネットワークにより単語列を出力し，画像キャプションを生成する．

画像キャプションのように，画像情報を別のドメインとして捉える手法としてシーングラフがある．シーングラフとは，画像中に含まれる物体をノードとしてそれらの関係性をエッジで表現し，画像情報を構造的に表現する手法である．また，シーングラフの生成を行なう研究分野がシーングラフ生成である．シーングラフ生成の分野では，多くの研究で画像情報のみからシーングラフが生成される．画像特徴量を元に，画像内の物体に付与された属性や物体同士の関連性を推測することでシーングラフを生成する．画像キャプションと比較すると，画像情報をテキストとして出力していない分，画像内容を理解できていないように感じるかもしれない．しかし，物体同士の関係性をグラフとして表現することで，テキスト表現よりも画像内容を明確に表現することができると考えている．

現在のシーングラフ生成では，上でも述べたように画像情報から構造を理解しようとする研究がほとんどである．しかし画像から得られた情報を全てグラフ化するため，シーングラフは膨大になってしまい，その画像における重要な要素が何かわからなくなってしまう．膨大なシーングラフの例を図 1.1 に示す．

またシーングラフの利用法として，画像の意味構造理解や画像生成などが挙げられるが，これらも複雑なシーングラフでは効果的に利用することができない．そこで，既存のシーングラフと画像キャプションを用いることで，より精度の高いシーングラフを生成することを目指した．精度が高いというのは，より簡潔なシーングラフで画像の意味構造を十分に表現することができることを表す．

これを示すため，二つの手法を提案する．まず，画像の本質的な情報抽出である．シーングラフにおいて，他の画像でも多く出現している部分グラフは，特定の画像の特徴を表して



$S = 1$ $S = 10$ $S = all$

図 1.1: 画像に含まれるシーングラフの部分情報の例を示す．画像に記してある S は，各画像に表示してあるシーングラフの部分情報の数である．

いないと考えられる．そこで，訓練データセット中で一定以上の出現確率に達した部分グラフを削除することで，より本質的なシーングラフが生成できると考え，評価実験を行なった．次に，複数出現する部分グラフの集約である．また，同一画像内で複数回出現する部分グラフは冗長に表現されており，ここをより適切な部分グラフに置き換えることで簡潔な表現にすることができると考えた．

シーングラフの改良精度を示す上で，新たな評価手法を提案する．シーングラフは，一般的な有向グラフや無向グラフよりも情報量が多い．そのため一般的なグラフ同型性判定問題で解くことは難しいので，シーングラフの改良精度を確認するための新たな評価指標を導入した．

実験の結果，画像の本質的な情報を抽出する実験に関しては，効果的な結果を示すことが出来た．多数の画像に当てはまる部分グラフを削除することで，各シーングラフには画像の特徴をよく示した本質的な情報が残ったと考えられる．同一シーングラフ内に複数回出現する部分グラフの変形については，精度に差が見られた．上手く変形できたところもあるが，あまり期待した置き換えができないことの方が多かった．これはキャプションに含まれている情報量の差が起因していると考えられる．評価実験の結果，全体を通してみるとシーングラフは改良されていることを示すことができた．

本稿の構成としては第 2 章で関連研究について述べ，第 3 章では本研究で使うシーングラフについて説明する．第 4 章で提案手法について述べ，第 5 章で提案手法を示すために行なった評価実験について述べる．第 6 章では評価実験の結果を考察し，第 7 章でまとめとする．

第2章　関連研究

本章では，本研究の関連研究について述べる．近年多く行なわれているコンピュータビジョン分野の研究の中でも，シーングラフと画像キャプションに焦点を当て説明する．これらは，どれも画像内容を構造的に表現し，人間にとって理解しやすくする試みである．シーングラフ表現，画像キャプション生成，シーングラフ生成の順に述べる．

2.1　シーングラフ表現

シーングラフとは，画像の意味情報を構造化した表現のひとつである．画像の中の物体情報，物体に付与されている属性情報，物体ペア間の関係性情報をグラフ構造で表現する．シーングラフとして表現することで，画像中の物体同士の関連性を明示的に捉えることができる．

シーングラフは的確に画像の意味情報を表現できることから，多くの研究分野に利用されている．例えば，代表的なものとして画像質問応答 [5] や画像検索 [6]，そして画像生成 [7] などが挙げられる．

Teney ら [5] は，VQA(Visual Question Answering) に対してシーングラフを効果的に利用している．VQA とは，画像内容に関する自然言語テキストの質問文に対して，画像情報を用いることで解答を生成する研究分野である．Teney らは，質問文の単語のつながりと画像の各物体をグラフ構造として捉え，それぞれの対応関係を学習させることで効果的な結果を生んだ．

また，画像検索の分野においては Jhonson ら [6] が，シーングラフを用いることで検索結果の向上が可能なことを示している．一般的に画像検索をする際には，求めている画像の特徴を的確に示すことで有用な結果を得ることができる．しかし，できるだけ特徴量を含めようと文章などの複雑な情報をクエリとしてしまうと，求めている画像とはかけ離れたものまで含まれてしまう．そこで Jhonson らは，画像情報をシーングラフとして捉えることで，複雑なクエリに対しても的確な結果を示すことが出来るのではないかと考え，結果として有用な精度を示している．

画像生成の分野では，Jhonson ら [7] がシーングラフを用いている．自然言語情報からの画像生成研究は存在するが，StackGAN[8] などによる画像情報を説明したテキストからの画像生成が一般的である．しかし Jhonson らは，シーングラフから画像を生成することで，より高精度かつ現実的な画像を生成できることを示した．このことから，テキストよりもシーングラフの方が画像情報を的確に表現することができると言える．

2.2　画像キャプション生成

画像キャプションとは，シーングラフと同様に画像の意味情報を表現する方法のひとつである．画像キャプションでは画像内容を自然言語のテキストとして表現する．画像の意味情報を表現する手段としては画像キャプション生成が一般的であり，シーングラフより

も多くの研究が行なわれている．そのため，画像の意味情報を理解する手段としての画像キャプションには触れる必要があると考え，本節では画像キャプション生成に関する研究について述べる．

画像キャプション生成に関する研究には長い歴史があり，ここ数年では驚異的な進歩を見せている．進歩を支える要因として挙げられるのが深層学習である．それらの研究において提案されている手法では，一般的に CNN(Convolutional Neural Network)[9] を使って画像特徴を抽出し，RNN(Reccurent Neural Network)[10] を使ってテキストを生成する．CNN も RNN も深層学習を用いた手法であり，画像認識や自然言語処理に関する以前の手法と比較して高い精度を示している．

Vinyals ら [11] による研究では，画像キャプション生成の代表的な手法である CNN と RNN を使用した手法を用いている．CNN で抽出した画像特徴を RNN に入力することで，画像情報を自然言語のテキストとして出力する．

CNN で画像特徴を抽出する際には画像を均一のグリッドに分けるのが一般的だが，Anderson ら [12] は，物体に焦点を当て特徴抽出をすることを提案した．物体などの特徴的な領域に焦点を当てることで，Attention[13] と呼ばれる計算機構が効果的に働く．これは，物体に着目してそれらの関係性を構造的に表すシーングラフと似ている．Anderson らの研究では個々の領域ごとのラベルは利用していないが，これらのラベルを利用することでシーングラフは作成される．

また近年では，教師なし学習による画像キャプション生成の研究 [14] もされている．Feng らの研究では，画像とキャプションのペアデータセットを必要とせず，GAN(Generative Adversarial Network)[15] のように生成器と識別器を利用してキャプションを生成する．教師なし学習では，キャプションを生成する上で画像全体よりも特徴的な物体の情報が重要となるため，精度の高いキャプションを生成するには物体同士の関連性を意識することが重要となる．

2.3　シーングラフ生成

画像キャプションと同様に，シーングラフを生成する手法も発達してきている [16, 17, 18]．しかし画像キャプション生成と比較すると量は少なく，まだまだ発展の可能性が残っていると考えられる．

物体認識の場合，状況が似ている 2 枚の画像を見分けることは難しい．例えば，「人」と「馬」が横に並んでいる画像が 2 枚あるとする．一方は人が馬に乗ろうとしているところで，もう一方は人が馬に餌を食べさせているところだったとしても，物体認識の結果としては「人」と「馬」という情報しか得ることが出来ない．そこで Xu ら [16] は，CNN と RPN(Region Proposal Network)[3] によって得られた領域情報からグラフ構造を推論することで，物体間の関係をより明示的に示すことができると考えた．

シーングラフの生成において重要なのは，正確に物体を認識することと，物体同士の関係性を推測することである．Li ら [17] は画像特徴を抽出する際に，領域特徴，フレーズ特徴，物体特徴をそれぞれ抽出することで高精度なシーングラフ生成を目指した．MSDN(Multi-level Scene Description Network) と呼ばれる手法で，様々な領域サイズで画像特徴を抽出することによって，より画像を理解することができるということを示した．

よりグラフ構造に焦点を当てたのは Yang ら [18] の研究である．Yang らはシーングラフを作成する際に，全ての認識した物体をグラフのノードとし，それら全てに対してエッジを結んだ．そこから物体同士の関連度を学習し，最終的に関連性の高いノード同士を結んだシーングラフが生成される．Yang らは関連度を求め，その結果に応じてエッジのみを切り

取っていった．しかし，物体として認識されたとしても，画像情報を的確に表すのに必要のないノードもあると考えられる．そこで，物体同士の関連度の低いエッジだけではなく，画像に対しての関連度の低いノードも切り取ることによって，より的確に画像情報を表すことを目指す．

第3章　シーングラフ

本章では，本研究において使用するシーングラフについて述べる．シーングラフの特徴とともに，本研究で提案するシーングラフの評価手法についても説明する．

3.1　シーングラフとは

本研究では，画像の意味情報を構造的に理解するためにシーングラフを用いる．本研究におけるシーングラフでは，物体情報を Objects，物体の属性情報を Attributes，物体ペア間の関係性を Relationships で表す．

シーングラフの例を図 3.1 に示す．図の右側に示されているのが，左の画像に対応するシーングラフの一例である．この例では，Objects として「woman」「bus」「truck」「bicycle」が示されている．「woman」を例として見ると Attributes は，衣服の色である「green」や「black」が付与されている．また Objects 間の Relationships は，「woman」と「bicycle」を繋げている「riding」などが挙げられる．

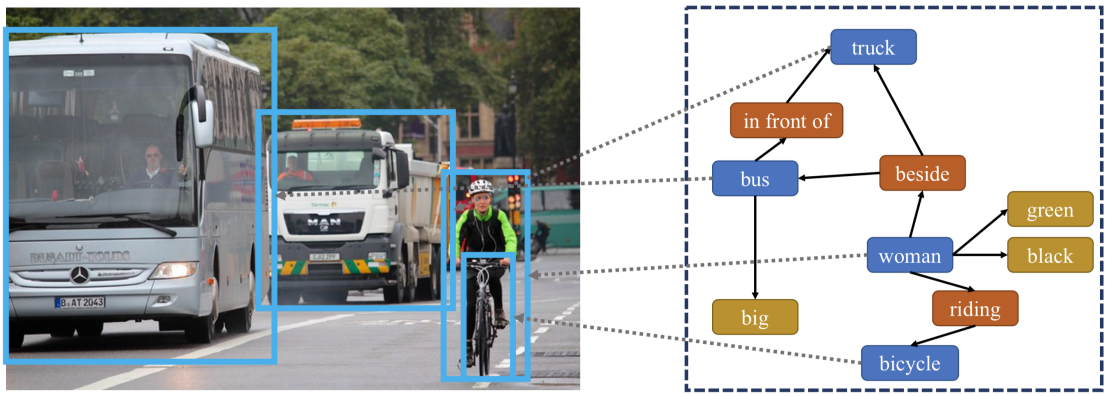


図 3.1: 左が元々の画像であり，右が画像から作成されたシーングラフの一例である．画像から物体を認識し，それぞれの関係性と物体の属性をグラフ構造で表現している．図のシーングラフでは，認識された物体情報である Objects を青色，物体の属性情報である Attributes を黄色，物体同士の関係性である Relationships を赤色で示している．

本研究におけるシーングラフの定義を示す．本稿では，シーングラフを $G = (O, E)$ で表す．O はグラフのノードを表し，E はエッジを表す．ここで O とは Objects の集合であり，$O = \{o_1, o_2, ..., o_n\}$ となる．また，E とは Objects のペア同士を結ぶ Relationships の集合であり，Relationships は $Objects - Relationships - Objects$ の形を取る．この Objects 間の Relationships をわかりやすくするため，$Objects - Relationships - Objects$ の関係性を

$head - edge - tail$ と表現する [19]．つまりシーングラフ G において，O は $head$ と $tail$ の集合であり，E は $edge$ の集合であると言える．

3.2　シーングラフの特徴

本節では，シーングラフでの表現の特徴を述べる．

まずシーングラフの大きな特徴は，名前の示す通り画像内容をグラフ構造で表すという点である．図 3.2 では，図 3.1 の左に示した画像に対するシーングラフの一部と，シーングラフに示されている部分のみのキャプションを示している．それぞれの Objects，Attributes，Relationships の色は，シーングラフとキャプションで対応している．

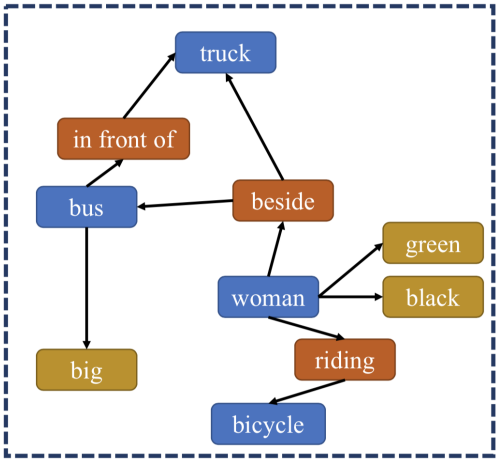


図 3.2: 同じ画像に対するシーングラフとキャプションの比較を示す．左側がシーングラフで，右側がキャプションである．シーングラフは一部だけを取り出し，キャプションはシーングラフに対応している部分を抜き出している．Objects，Attributes，Relationships はそれぞれ青色，黄色，赤色で色付けされ，この色はシーングラフとキャプションで対応している．

図 3.2 を見ると，シーングラフよりキャプションの方が細かい情報まで示しているように感じられる．しかし，これはキャプションを文章で表現するという都合上，シーングラフ上には載せることが出来なかった情報も付加しているためである．実際には，情報量はさほど変わらず，表現方法のみの違いと言える．情報量が変わらないと言っても，キャプションは自然言語の文章で表現されるため，同じ主語が何度も現れたり，Objects 間の関係性がわかりにくかったりと問題点もある．対してシーングラフは，複雑な情報を簡潔な構造にすることで，キャプションより理解しやすいものとなっている．

特にシーングラフの特徴として挙げられる点は，画像理解の容易さである．図 3.2 を見てもらえるとわかるが，キャプションの方は詳細に記述されているにも関わらず，画像を思い浮かべると多様な情景が浮かぶのではないだろうか．対してシーングラフは，画像の意味構造を元にグラフとして表現されているため，画像内容を正確に記述することができるのである．つまり，画像からシーングラフやキャプションへの変換は同程度の情報量を残せたとしても，シーングラフやキャプションから画像内容への変換には大きな差が出てしまうと言える．この点が，シーングラフの一番の特徴として挙げられる．

3.3　シーングラフの評価手法

3.3.1　グラフ構造の評価

シーングラフを改良する上で精度の評価をする必要があるが，シーングラフの評価には定まった手法がない．シーングラフはグラフ構造ではあるものの，一般的な無向グラフや有向グラフよりも情報量が多い．有向グラフと同様にエッジにはあるノードからあるノードへの向きがあり，ノードに加えてエッジにもラベルが付与されている．これでは一般的なグラフ同型性判定問題などでの解決は難しい．

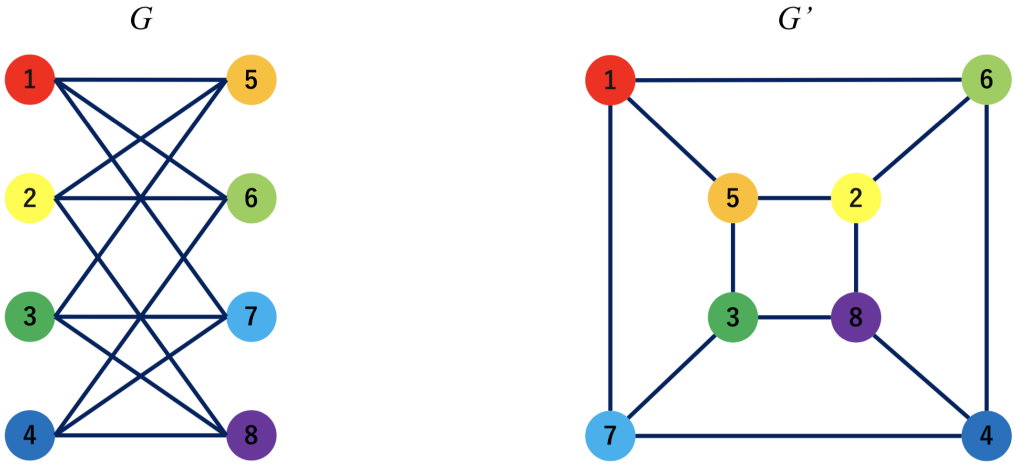


図 3.3: グラフ同型性判定問題の例を示す．このグラフ G とグラフ G' は同型であると言える．これらは同数のノードと同数のエッジを持ち，どれも同一のつながり方をしている．グラフ同型性を判定する問題はとても計算量が多く，解くのが難しい．

図 3.3 に示すように，グラフ同型性判定問題とは，二つのグラフが同じ形をしているかどうかを判定する問題のことである．この場合の同じ形というのは，2 つのグラフが同数のノードと同数のエッジを持ち，全てのノードが同一のつながり方をしている状態のことである．グラフの同型性を判定する問題はとても計算量が多く，現実的な時間で解くことは難しい．また，本研究では改良したシーングラフが元シーングラフと同型である必要はないため，別の評価手法を導入する．

3.3.2　本研究でのシーングラフ評価手法

本研究では，シーングラフの改良精度を評価するために，新たな評価手法を導入する．本研究の目的は，意味構造に着目することでシーングラフの精度を上げることである．シーングラフの精度を上げるとは，画像から得られる膨大な情報を全てグラフにしてしまうことから，画像の特徴をよく表している要素のみを簡潔にグラフにすることである．

評価手法としては，いくつかの要素を組み合わせた総合的な評価によって精度を求めることとする．具体的にはシーングラフのノードの数，エッジの数，あらかじめ選択した代表 Objects が過不足なく含まれているか，以上の 3 点によって評価する．あらかじめ選択した代表 Objects とは，画像内容を示す上で視覚的に重要だと思われる Objects 5 つと，画像には含まれているが本質的ではない Objects 5 つである．これらの Objects については，一般

的に重要だと考えられるものを，本稿では選択した．画像内容を示す上で視覚的に重要だと思われる Objects 5 つがどれだけシーングラフに含まれているかと，画像には含まれているが本質的ではない Objects 5 つがどれだけ含まれていないかによって評価する．これらは最低でも Objects が 10 以上ある画像でしか評価できないため，テストデータセット全てではなく，テストデータセットから選択した 100 の画像を使い評価する．また，正当な評価を行なうため，これらは評価実験の前にあらかじめ選択しておく．以下の図 3.4 に選択した Objects の例を示す．ノードの数とエッジの数は，どの程度シーングラフが本質的な情報かつ簡潔な表現になったかを評価するために用いる．

図 3.4: 評価のために選択した Objects の例を示す．*image* は対象の画像，*positive* は画像の特徴をよく表す Objects，*negative* は画像には含まれているが本質的ではない Objects である．

第4章　提案手法

4.1　概要

本研究では，シーングラフとキャプション情報を用いることで，より良いシーングラフを生成することを目指す．本章ではシーングラフを改良するための手法を提案する．4.2 では統計的な観点により画像の意味構造において重要なエッジを見極める．4.3 では冗長的な表現になってしまうシーングラフを，キャプションを用いることで簡潔な表現にする．これらを行なうことで，画像の意味構造を表現するのに簡潔かつ十分なシーングラフを生成できると考えた．

4.2　意味構造において重要でないエッジの削除

画像から得られる特徴には，画像が示している情報を的確に表している特徴と，画像の意味を考える上でそこまで重要ではない特徴があると言える．図 4.1 の通り，左の画像に示された領域と右に示された領域を比較すると，画像の意味構造を求める上では左の画像から得られた領域の方が重要であると考えられる．左の画像では画像中央にある「ソファ」を，右の画像では「ソファ」の後ろにある「ライト」に領域が示されている．この画像例の場合，重要な要素と考えられるのは「ソファ」「テーブル」「窓」「部屋」などが挙げられる．つまり右で示された領域は，実際に画像の意味構造を考える上で重要ではない要素であると考えられる．そこで，このような画像の意味構造において重要ではないと考えられる要素が，シーングラフに反映されないようにすることを目指した．

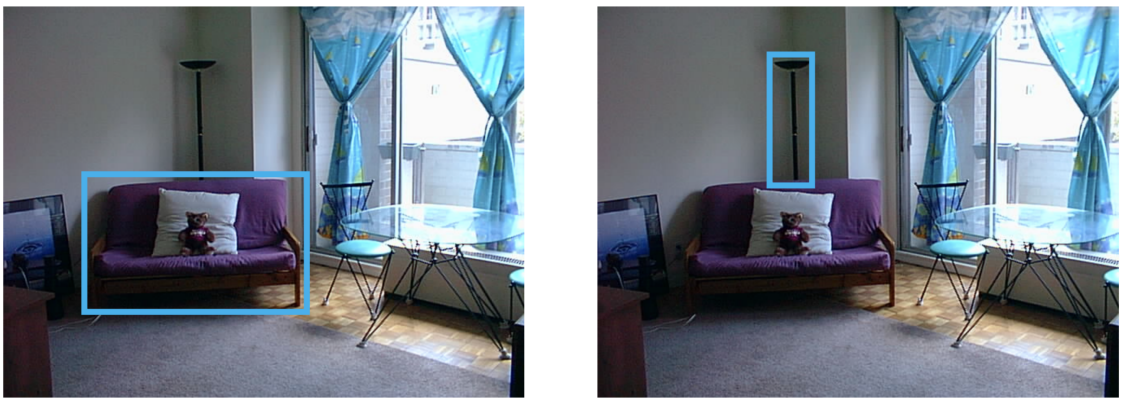


図 4.1: 意味構造における重要な画像領域の一例を示す．同じ画像に対してでも，焦点を当てている物体の違いによって，画像情報の意味構造は変化してくる．左は画像中央にある「ソファ」に領域が示されていて，右は後ろにある「ライト」に領域が示されている．この画像の意味構造を考えた時，重要となるのは左の「ソファ」であると考えられる．

このような画像の意味構造において，重要でないエッジを削除するにはどうすればよいだろうか．この問題を解決するには，画像から認識された Objects とそれらの Relationships において，多数の画像において出現頻度の高いものは重要ではないのではないかと考えられる．さらに，それらの出現確率による手法の分類を行なった．

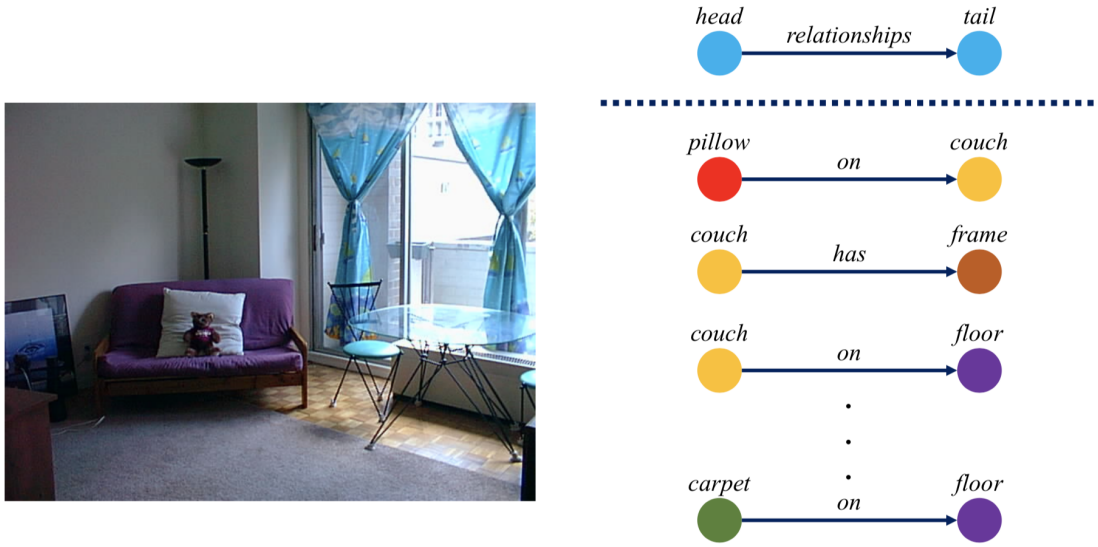


図 4.2: 画像から得られる $head - edge - tail$ 三つ組の一例．同一の Objects は同色で示し，Objects 間の右向き矢印は Relationships を表す．つまり Objects がノード，Relationships がエッジのグラフ構造となっている．

4.2.1 同時確率による重要度の低いエッジの削除

図 4.2 では，同一画像から得られる複数の $head - edge - tail$ の三つ組の一部を列挙している．この図を見るとわかる通り，$head - edge - tail$ はその画像の特徴を明確に表す三つ組もあれば，他の画像における同じような光景を表す三つ組も存在する．図 4.2 に示されている例で言えば，$pillow - on - couch$ は，ただ部屋にソファが置いてあるだけではなく，ソファの上には枕がある点を明示している．これは画像の特徴を表すと言ってよいだろう．しかし，反対に $couch - on - floor$ などは画像の特徴を示す上で，あまり重要ではないのではないかと考えられる．その理由としては，ソファの情報が他で示されていた場合ソファが床にあるのは当然として考えられる点や，ソファがある画像はたくさんあるので更に深い情報でないと意味がない点などが挙げられる．つまり今回の例で言うと，多数の画像が含まれたデータセットで考えた場合 $couch - on - floor$ は $pillow - on - couch$ より多く出現していると考えられる．

このように，$head{-}edge{-}tail$ の情報が全て同じだった場合のみの出現回数を，$(head,edge,tail)$ の同時確率による出現回数として捉えることができる．同時確率とは，複数の事象が同時に起きる確率のことである．X と Y が同時に起きる確率は，$P(X,Y)$ で示され，以下の式で計算される．

$$P(X,Y) = P(X)P(Y) \tag{4.1}$$

つまり，$head-edge-tail$ の全ての要素が同じ時の出現確率は，$P(head, edge, tail)$ として捉えることができる．

訓練データにおいて，同時確率に基づいて出現頻度の多い $head-edge-tail$ の三つ組をテストデータでは削除することで，重要な要素のみを取り出せると考えられる．三つ組を削除するという作業は，$head-edge-tail$ を全て削除するのではなく，$edge$ と $tail$ を削除する作業のことである．つまりシーングラフ上には $head$ のみが残り，これを繰り返すことで，再帰的に削除していく．削除を行なう条件式は以下の通りである．

$$\sum_{i \in I} \sum_{x \in E} \tau(i) P(x,t) > C \tag{4.2}$$

I はデータセットに含まれる画像の総数であり，$\tau(i)$ は一つの画像に含まれる E の総数である．E の出現回数をシーングラフごとに加算し，さらにデータセット中の全ての画像に対して加算することで，全ての画像に対する E の出現頻度を求める．そして，出現頻度が任意の値 C を越えた場合削除を行なう．

4.2.2 条件付き確率による重要度の高いノードの検出

4.2.1 では同時確率による出現頻度を元にエッジの削除を行なったが，本項では条件付き確率によるノードの検出を行なう．条件付き確率とは，ある事象 Y が起きた時に，その条件下で事象 X が起きる確率のことである．$P(X|Y)$ で示され，以下の式で計算される．

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{4.3}$$

4.2.1 では $head-edge-tail$ が全て同じ場合を同時確率として捉えた．同様にして，ある $head$ から $edge-tail$ が出現する場合や，$head-edge$ から $tail$ が出現する場合を条件付き確率として捉える．

同時確率による出現回数では，全てのシーングラフにおいて $head-edge-tail$ が全て同一の場合の出現回数を数える．本項では同じシーングラフ i において，$head$ が同一のものから $edge-tail$ が生じる場合と，$head-edge$ が同一のものから $tail$ が生じる場合を，それぞれ以下の式で求める．式 4.4 が，$head$ が同一のものから $edge-tail$ の三つ組が生じる場合であり，式 4.5 が，$head-edge$ が同一のものから $tail$ が生じる場合である．ここで S はシーングラフに含まれるエッジの総数である．

$$\sum_{x \in E} \tau(i) P((edge-tail)|head) \tag{4.4}$$

$$\sum_{x \in E} \tau(i) P(tail|(head-edge)) \tag{4.5}$$

データセット中の多数のシーングラフに複数回出現している要素は，画像の本質的な特徴を示さないと考えられる．反対に，個々のシーングラフにおいて複数回出現する要素は，その画像において重要な特徴を表していることが多いと考えられる．そこで，上記の式をシーングラフに適用し，その画像の意味構造において重要な要素を検出する．

4.2.1 では同時確率によって出現回数を求め，一定の値を超えた $head-edge-tail$ の三つ組は全て削除した．本項では，条件付き確率による出現頻度により各ノードの重要度を求め，重要度が低い部分グラフを削除することで，さらに精度を上げることを目指す．

4.3 複数出現する要素の検討

4.2 では統計的手法に着目し，多量のデータにおいて出現頻度の高い $head - edge - tail$ の三つ組は画像の特徴を表さないと考え，その三つ組を削除することで精度向上を目指した．また，ひとつのシーングラフにおいて複数回出現する要素は重要であると考え，重要度が低い部分グラフを削除することを提案した．本節では，ひとつのシーングラフで複数回出現する $head - edge - tail$ の三つ組の扱いについて検討を行なう．

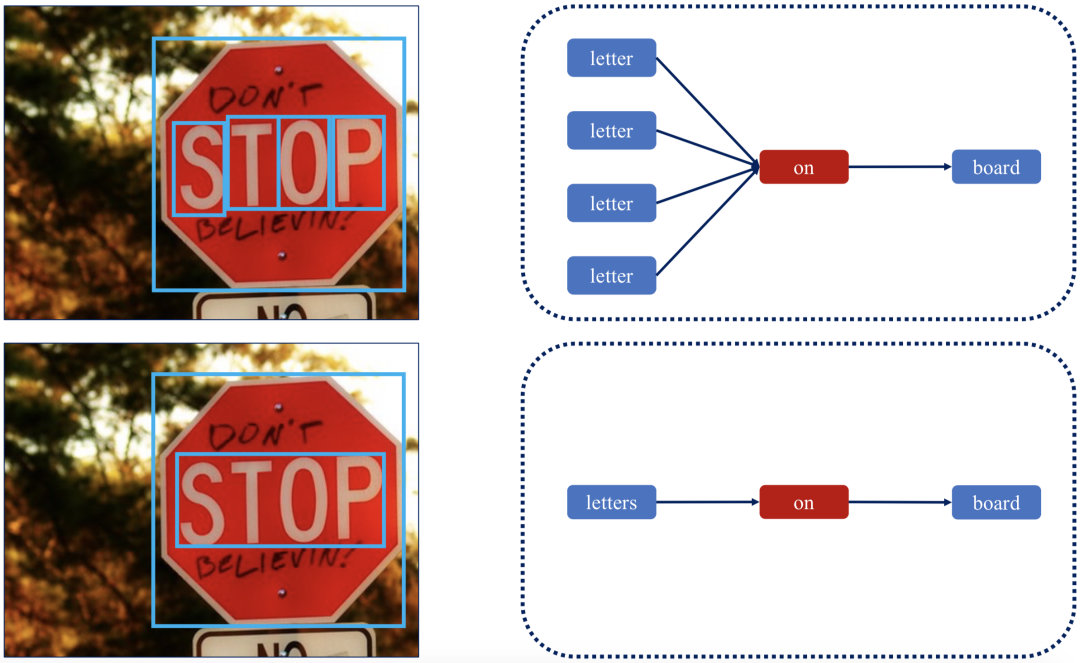


図 4.3: ひとつのシーングラフにおいて $head - edge - tail$ の三つ組が複数回出現する画像の例を示す．上が現状問題のある生成されたシーングラフである．本来は一文字ずつ認識するのではなく，文字列として認識することで意味を理解することができる．下のようなシーングラフになってくれるのが理想である．

図 4.3 で示すように，画像から生成されたシーングラフには，同じ $head - edge - tail$ の三つ組が複数回出現することがある．この現象は図のような文字が書いてある場合だけではなく，たくさんの建物が写っている画像などの場合にも生じることがある．しかし図を見るとわかるように，これでは冗長性であるため，図 4.3 の下のようになるのが理想的であると考えられる．

そこで，このような課題を解決するために，複数出現した $head - edge - tail$ の三つ組をまとめることでシーングラフを簡潔にすることができると推測した．ひとつのシーングラフで $head - edge - tail$ の三つ組が複数回出現した場合，それらをまとめ，$head$ を変形することで簡潔なシーングラフにする．図 4.3 の例で示すと，$letter - on - board$ が複数回出現しているので，$letters - on - board$ のように $letter$ を複数形にし，まとめることで簡潔な表現にする．$head$ の変形は Word2Vec の単語類似度とレーベンシュタイン距離によって行なうため，まず Word2Vec[20] とレーベンシュタイン距離について説明する．続けて複数出現する部分グラフの削除手法について説明する．

4.3.1 Word2Vec とレーベンシュタイン距離による単語の類似度

まず，Word2Vec[20] について説明する．Word2Vec とは，簡単に言うと単語をベクトルとして表現することで次元を圧縮する手法のことである．ベクトルとして表現することで，単語の意味や文法などを捉えることができるようになる．Word2Vec では，ベクトル化する際に分散表現を使うことで次元を圧縮することができる．

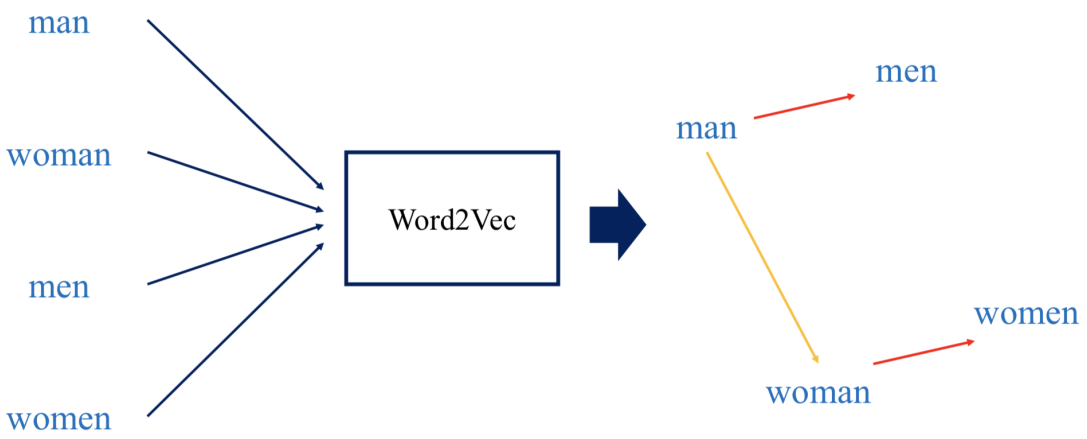


図 4.4: Word2Vec のイメージを表した図を示す．左が単語の入力例であり，右が入力された単語が埋め込まれたベクトル空間のイメージである．*man* と *men*，*woman* と *women* はそれぞれ同一の方向性を持ち，距離も近い．また同様に *man* と *woman*，*men* と *women* も同一の方向性を持っているが，こちらは上のペアよりも関係性が遠いことから，距離を離して表現している．これはあくまでイメージであり，実際のベクトル空間において，図のように埋め込まれている訳ではない．

図 4.4 に Word2Vec のイメージを載せる．一番左側に示しているのが，入力された単語の例である．単語列を Word2Vec のニューラルネットワークに入力すると，一番右側に示しているように同じベクトル空間に埋め込まれる．図 4.4 はイメージであるが，同じ空間に埋められた単語は，似た意味や文法によって同じ方向性や近い距離に表れる．図に示した単語例を用いて説明すると，*man* と *men*，*woman* と *women* はそれぞれ同一の方向性を持ち，距離も近い．また同様に，*man* と *woman*，*men* と *women* も同一の方向性を持っている．しかし，こちらは先ほどの関係よりも距離が離れていて，これは先ほどよりも関係性が遠いことを表している．これを用いることで，画像に対応するキャプションから最適な類似単語を見つけることができる．

しかし Word2Vec による類似単語では，その性質上，全く逆の意味でも類似度が高くなってしまうという問題がある．そこで Word2Vec に加え，レーベンシュタイン距離を測ることで問題を解決する．

レーベンシュタイン距離とは単語同士の類似度を測る数値であり，ある文字列から目的の文字列への編集距離のことである．編集距離とは，ある文字列から目的の文字列に変形するまでの最小変形回数のことである．Word2Vec とレーベンシュタイン距離を組み合わせ，最終的な類似単語を見つける．

4.3.2 複数出現する部分グラフの集約

複数出現する部分グラフの削除には，上で説明した Word2Vec とレーベンシュタイン距離を利用する．まず各シーングラフに対して $head - edge - tail$ の同時出現数を以下の式 4.6 で求め，一定回数以上出現した場合にまとめる．

$$\sum_{x \in E} \tau(i) P(x,t) > C \tag{4.6}$$

ここで C は任意の値とする．各シーングラフ i の中で C 以上の出現回数であった $head - edge - tail$ の三つ組に対して，部分グラフをまとめる．式 4.7 を満たす場合 $head$ を変形し，満たさない場合は $head$ は変形せず，部分グラフのまとめだけ行なう．式 4.7 では Word2Vec とレーベンシュタイン距離を組み合わせ，最適な類似単語への変形を行なっている．

ここで，LD とはレーベンシュタイン距離を求める関数であり，WV とは Word2Vec での類似度を求める関数である．また，x には任意の $head$ が当てはまり，t には類似度を求める対象である同一シーングラフ内の Objects が当てはまる．

$$(LD(x,t) \leq 3) \wedge (WV(x,t) > 0.4) \tag{4.7}$$

その後，まとめた部分グラフを元の場所に置換する．必要な場合これを再帰的に行ない，簡潔なシーングラフを生成する．

第5章　評価実験

5.1　概要

本章では，提案手法を評価する実験について述べる．評価実験では，実際に提案手法に基づいてシーングラフを生成し，ベースラインと比較する．まず評価実験に用いるデータセットと，それらに行なう前処理について説明する．次に，提案手法を評価するために行なった実験の内容について述べる．最後に実験結果についてまとめる．

5.2　データセット

本研究では，データセットとして Krishna らによって作成された Visual Genome[21] と Lin らによって作成された MSCOCO(Microsoft Common Objects in Context)[22] を使用する．それぞれのデータセットの特徴について述べ，実験前に行なう前処理について説明する．

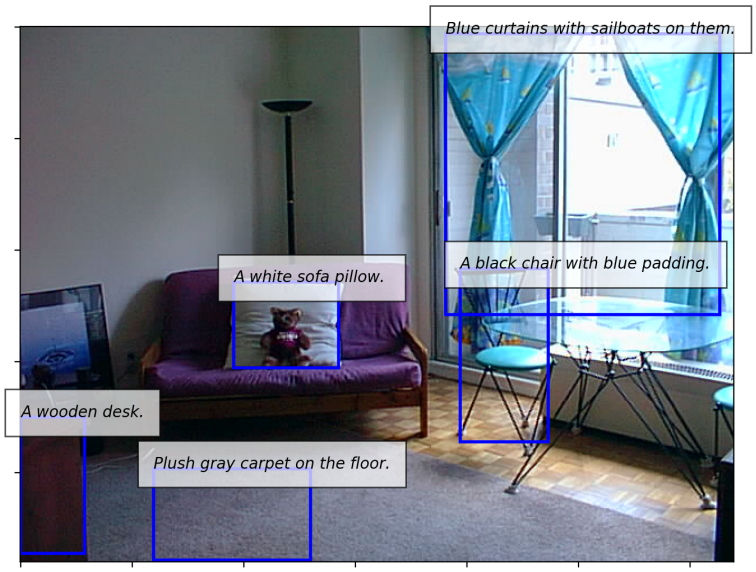


図 5.1: Visual Genome データセットに含まれる画像の一例を示す．物体情報が矩形で示され，矩形で囲まれた小領域毎にキャプションが付与されている．見やすさのため，データセットで示された矩形のうち 5 つを表示している．

5.2.1 Visual Genome

Visual Genome[21] は，108,077 枚の画像が含まれており，各画像にシーングラフが付与されているデータセットである．つまり各画像ごとに，画像中に含まれる物体 (Objects)，物体に付属する属性 (Attributes)，2 つの物体ペア間の関係性 (Relationships) が付与されている．Objects には各画像中に含まれる物体名と物体領域が，Attributes には各物体毎に色や形状の属性情報が，Relationships には物体ペア間の関係性情報が含まれる．また図 5.1 に示す通り，特定の小領域に対してキャプションが付与されているが，画像全体のキャプションは付与されていない．Visual Genome は，シーングラフを用いる用いないに関わらず，多くの研究 [5, 16, 17, 18, 23] で利用されているデータセットである．

図 5.2: MSCOCO データセットに含まれる画像の一例を示す．左が元画像で，右は物体領域が示された画像である．画像の特徴となるような大きい物体だけではなく，画像右下の小さいボトルにも領域が示されている．また物体領域だけではなく，物体名も詳細に記述されている．画像の下にあるテキストは画像に対するキャプションであり，各画像に対して平均 5 つ付与されている．

5.2.2 MSCOCO(Microsoft Common Objects in Context)

MSCOCO[22] は Microsoft によって提供され，約 330,000 枚の画像を含み，各画像にキャプションが付いているデータセットである．各画像につき平均 5 つのキャプションが付与され，物体名と物体位置が詳細に示されている．また図 5.2 に示す通り，画像に対して物体の領域が示され，画像右下のボトルのような小さい物体にも対応している．そのため，このデータセットは多くの画像認識 [24] や画像キャプション生成 [14, 25] の研究に利用されている．

5.2.3 前処理

評価実験で使用するデータセットの前処理について説明する．各画像に対して，シーングラフとキャプションが付与されているデータセットを使用するため，Visual Genome と MSCOCO を加工して作成する．

Visual Genome には各画像に対して Objects，Attributes，Relationships が含まれたシーングラフが付与されているが，画像全体に対してのキャプションは付与されていない．反対に，MSCOCO には複数の詳細なキャプションが付与されているが，シーングラフに含まれる情報はない．これらの問題を解決するため，2 つのデータセットを用いて新たなデータセットを作成する．

Visual Genome 内のいくつかの画像は，Flickr[26] と MSCOCO から収集されている．そこで Visual Genome と MSCOCO に共通する画像のみを使用することとした．共通の画像を収集した結果，51,498 枚の画像を収集することができた．これらの画像には，シーングラフと平均 5 つのキャプションが付与されている．前処理の結果を以下の表 5.1 に示す．

表 5.1: 前処理後のデータセットの内容

	全て	訓練	テスト
画像枚数	51,498	36,048	15,450
シーングラフの総数	51,498	36,048	15,450
ノードの総数	1,813,241	1,269,347	543,894
エッジの総数	1,110,813	777,126	333,687
ユニークなノードの総数	-	48,879	27,427
ユニークなエッジの総数	-	265,780	132,799

表 5.1 に示す通り，訓練画像とテスト画像については，全 51,498 枚のうち 7 割の 36,048 枚を訓練画像，3 割の 15,450 枚をテスト画像とした．以降は，訓練画像で実験を行ない，テスト画像で評価を行なう．また画像毎にひとつのシーングラフが付与されていることから，シーングラフの内訳も画像枚数と同等である．表の「ノードの総数」に示しているのは Visual Genome に含まれている Objects と Attributes の情報であり，値を見ると画像 1 枚につき約 35 のノード情報が付与されていることがわかる．同様に，表の「エッジの総数」を見ると各画像につき約 21 のエッジ情報が付与されていることがわかる．また，ユニークなノードの総数とユニークなエッジの総数というのは，それぞれの重複するノードとエッジを削除した総数のことである．

5.3 本質的な意味構造の抽出

前処理を行なったデータセットを元にして，本質的な意味構造抽出するための実験を行なう．本節では，4.3 で示した手法に基づいて実験を行ない，実験の結果を示す．

前処理されたデータセットにおいて，シーングラフ $G = (O, E)$ における E の出現頻度を確認する．この作業は訓練データセットにおいてのみ行ない，結果を考慮してテストデータセットに反映させる．以下に，式 4.2 を訓練データセット 36,048 枚に対して行なった結果を示す．

まず，式 4.2 において N は訓練画像の総数である 36,048 となり，S は個々のシーングラフ毎の $head - edge - tail$ の三つ組の数である．実際に $head - edge - tail$ の出現頻度を計算した結果，出現回数が 5 回以下の $head - edge - tail$ は 247,931 で訓練データに含まれる

表 5.2: 出現回数が多い $head - edge - tail$ の例

$head - edge - tail$	出現回数
clouds in sky	4,438
window on building	3,536
man wearing shirt	3,107
sign on pole	1,144
window on train	1,144

ユニークなエッジ数の 93.3 %であった．訓練データの中で一番多かった「*clouds in sky*」の出現回数は 4,438 回であり，訓練データの中でも $head - edge - tail$ の出現回数には大きな差があることがわかる．

表 5.2 に出現回数が多い $head - edge - tail$ の例を示した．表からわかるように，明らかに出現回数の多い $head - edge - tail$ は，画像中の背景に当たるような，画像内容にあまり関係がないと考えられるものが多い．そのため私は，ある一定の値を超えた出現回数を持つ $head - edge - tail$ は，画像の本質的な意味構造を抽出する上で必要がないと考えた．

出現回数が 5 回以下の $head - edge - tail$ がデータ中の 93.3 %を占めるという結果から，今回の評価実験では出現回数 5 回をベースとして実験を行なうこととした．出現回数を調べた結果，出現回数 10 回以下で 96.8 %，100 回以下で 99.8 %となったため，評価実験では出現回数 6 回以上のものを削除することで十分だと判断した．そこで，今回の場合では式 4.2 の C は 6 とする．

さらに同時確率による出現頻度だけではなく，条件付き確率による出現頻度についても触れる．$head - edge - tail$ の全てが同一の場合のみ削除を行なおうとしたが，これではひとつのシーングラフ内で 6 回以上出現した三つ組や，少数の画像で複数回出現した三つ組を削除してしまうことになる．しかし，それらは果たして本質的でないと言えるだろうか．むしろ画像の特徴を表している可能性も大きいと考えられる．そこで，式 4.4 と式 4.5 に基づいて，各シーングラフ内のノードとエッジの出現頻度を求め，エッジ接続数が 4 以上のノードは削除を行なわなかった．

本節で行なう評価実験では，上で示した削除すべき対象をテストデータセットに対して適用し，シーングラフの変化を確かめる．元のシーングラフと手法適用後のシーングラフの比較の一部を図 5.3 に示す．全てのノードとエッジを載せることは難しいので，変化が起きた部分のグラフを示す．

図 5.3 ではシーングラフの一部しか載せることができていないが，ここで示したものを見るだけでも，無駄なエッジが削除できていることがわかる．もちろん全てが上手くいっているわけではないが，実際に結果を見ると効果的に作用していると考えられる．

5.4 複数出現する要素の評価

次に，複数回出現するエッジの検討を行なう．4.3 で示した提案手法に基づいて，複数回出現する部分グラフについての実験をし，結果を評価する．

式 4.6 において C の値を 2 とし，以下の実験を行なう．まずテストデータセットに対して，シーングラフ内に複数出現する $head - edge - tail$ の三つ組の出現回数を求め，それらに対して提案手法を適用する．部分グラフの $head$ に対して，Word2Vec とレーベンシュタイン距離を組み合わせ，対応するキャプションから最も類似度の高い単語を抜き出す．Word2Vec の学習は，全てのデータセット中の画像に対応するキャプションを用いて行なった．

図 5.4 では，図 5.3 と同様に結果の一部を示す．図を見てもわかる通り，上手く置換出来たものと出来なかったものが存在した．結果を確認したところ，*boy* や *girl* などの単語は上手くいかず，*book* や *newspaper* などは比較的高い精度を示していた．*boy* や *girl* は，それぞれの複数形である *boys* や *girls* よりも，*kid* や *child* などの方が類似度が高くなっていた．しかし全体を通して見ると，精度が高いとは言えなかった．これは Word2Vec による類似度が原因であると考えられる．

Word2Vec では，単語を同じベクトル空間に埋め込み，ベクトル同士の距離や方向によって類似度を求める．ベクトル空間への埋め込みは，訓練データセットに含まれるキャプションを用いて行なっているため，キャプション上で用法が似ている単語は類似度が高く出力される．また，キャプション上に複数形として存在しない場合や適切な単語が存在しない場合，上手く変形することが出来なかった．そのため，置き換えても自然となるような単語は上手くいかず，置き換えた単語よりも複数形の方が自然となる単語は上手くいったと考えられる．

5.5 手法の比較結果

上記で示した評価実験の結果について述べる．3.3 で示した評価手法を元に，選択した 100 の画像について結果を比較し，以下の表 5.3 に示す．元のシーングラフを SG，本質的な意味構造の抽出を行なったシーングラフを $SG-E$，複数出現した部分グラフの検討を行なったシーングラフを $SG-M$，どちらも適用したシーングラフを $SG-E-M$ とする．positive Objects 数は多いほど良く，negative Objects 数は少ないほど良い結果となる．元のシーングラフは全ての Objects を含んでいるため，どちらも最大値の 5 となっている．

表 5.3: 評価実験結果

	平均ノード数	平均エッジ数	平均 positive Objects 数 (Max:5,Min:0)	平均 negative Objects 数 (Max:5,Min:0)
SG	35	21	5	5
SG-E	26.3	14.5	4.7	2.0
SG-M	30.1	18.2	**4.9**	3.8
SG-E-M	**22.1**	**12.2**	4.7	**1.4**

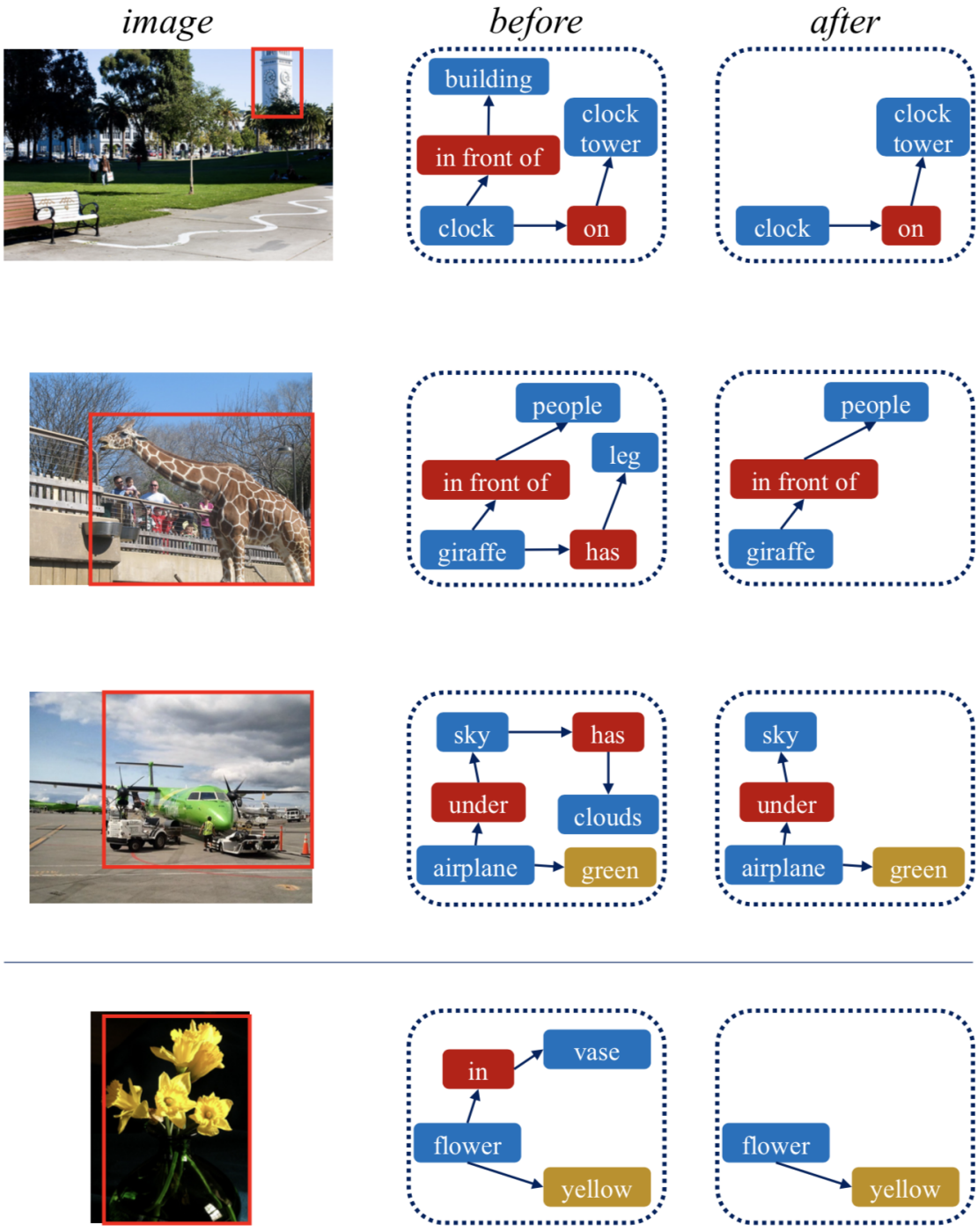


図 5.3: 元のシーングラフと，新たに生成したシーングラフを比較した．Objects を青色，Attributes を黄色，Relationships を赤色で示している．線を境に，上に示した 3 つは上手く改善されたシーングラフであり，一番下は本質的な情報まで削除されてしまったシーングラフである．上の 3 つはそれぞれ画像の本質ではない部分グラフが削除されているが，一番下は画像における重要な要素が失われてしまっているように感じられる．手法適用後のシーングラフでは，花があることがわかるが，ただ花があるだけで花瓶に入っているかがわからない．

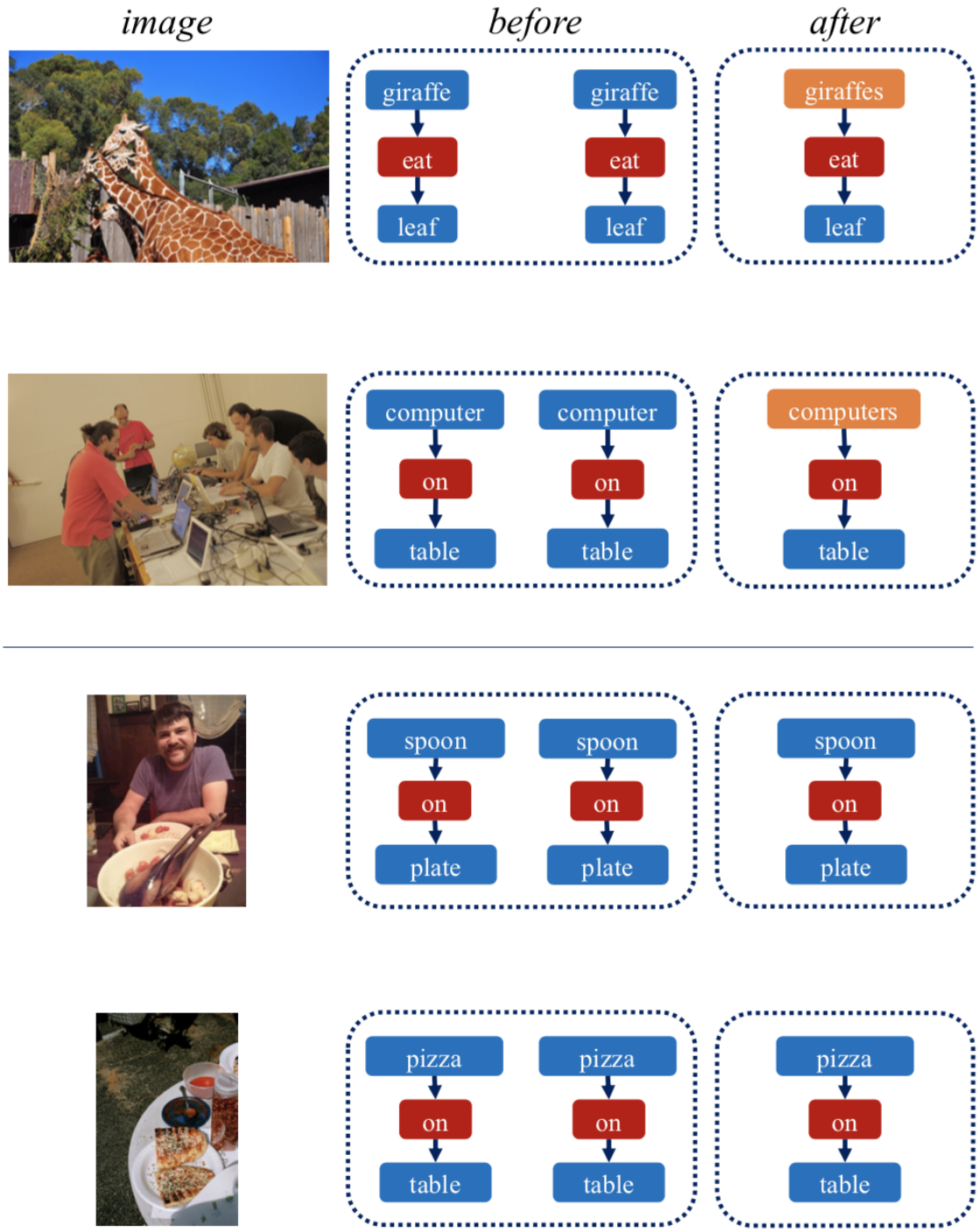


図 5.4: 複数出現する要素のまとめ実験により生成されたシーングラフの比較を示す．左列が元の画像，中央列が手法適用前のシーングラフの一部，右列が手法適用後のシーングラフの一部である．中央列のシーングラフは，本来図で示しているものと同様の三つ組が複数あるが，ここでは見やすさのためひとつのみ載せている．Objects を青色，Relationships を赤色で示し，変更が行なわれたノードはわかりやすさのため橙色とした．

第6章　考察

評価実験を行なった結果，表 5.3 を見るとわかる通り SG-E-M で精度向上を示すことができた．出現頻度に基づくエッジの削除においては，複雑な画像に対しては特に有用性を示すことができたが，少ない物体で構成された単純な画像では，精度向上があまり見られなかった．これは，他の画像では本質的でない部分が，主要の要素となっていることから起きてしまったと考えられる．しかし，元々シーングラフは複雑な画像内容を，物体同士のグラフ構造で表すことで理解しやすくするという特徴があるため，このことは問題ではないと考えている．また，同時確率だけではなく条件付き確率を用いることにより，削除するべきではない三つ組を残すことができた結果が，いくつかのサンプルで見られた．

同一のシーングラフ内で複数出現する部分グラフの検討については，単体ではあまり有用性を示すことができなかった．原因としては，キャプションに含まれる語彙の情報量によって精度に差ができてしまったためだと考えられる．キャプション上から類似単語を見つけ出すため，キャプション上に存在しない単語には置き換えることができない．そのため，類似単語候補が閾値を超えず，ただまとめるだけのグラフが多かった．しかし，複数出現している部分グラフをまとめることで，簡潔なシーングラフとなることは示すことができた．また nagative Objects の削減にも効果を示していた．

上記の 2 点を組み合わせた結果，簡潔かつ十分な情報量のシーングラフが生成できることを示すことができた．

第7章　おわりに

本研究では，意味構造に基づいたシーングラフ生成手法の提案を行なった．シーングラフの説明から入り，画像キャプションとの比較を行ない，現在のシーングラフ生成は画像情報を詳細に示そうとしすぎることで，シーングラフが巨大になり，視覚的に理解しづらくなってしまうという課題点を示した．そこで，元のシーングラフと統計情報を用いることで，画像の内容をより正確に示せるのではないかと推測した．また，冗長な部分グラフを正すことで，グラフ情報を簡潔にできるのではないかと推測した．

2 つの課題点を解決するために，それぞれの提案手法を示した．ひとつは，訓練データセットの統計情報を元にして，不要な部分グラフを削除するという手法である．もうひとつは，単語の類似性を用いることで，複数回出てくる部分グラフの情報をまとめるという手法である．提案手法を示すための評価実験を行なった結果を示し，有用性を示すことができた．

今後の展望としては，シーングラフからの画像生成などが挙げられる．本研究で提案した手法を用いることで，より画像内容に近いシーングラフを生成することができた．このことより，一般的なシーングラフよりも簡潔な表現となり，本質的な画像を生成することができると考えられる．

謝辞

本研究を進めるにあたり，数々のご指導を頂きました筑波大学図書館情報メディア系手塚太郎准教授に心より感謝いたします．また，辛い時に支えになってくださった全ての皆様に感謝いたします．

参考文献

[1] S.Ren, K.He, R.Girshick, and J.Sun. Fater r-cnn: towards real-time object detection with region proposal networks. In Neural Information Processing Systems(NIPS), 2015.

[2] Q.Zhao, T.Sheng, Y.Wang, Z.Tang, Y.Chen, L.Cai, and H.Ling. M2det: a single-shot object detector based on multi-level feature pyramid network. Association for the Advancement of Artificial Intelligence(AAAI), 2019.

[3] S.Ren, K.He, R.Girshick, and J.Sun. R-cnn: towards real-time object detection with region proposal networks. In Neural Information Processing Systems(NIPS), 2015.

[4] J.Redmon, S.Divvala, R.Girshick, and A.Farhadi. You only look once: unified, real-time object detection. In Computer Vision and Pattern Recognition(CVPR), 2016.

[5] D.Teney, L.Liu, and A.van den Hengel. Graph-structured representations for visual question answering. In Computer Vision and Pattern Recognition(CVPR), 2017.

[6] J.Jhonson, R.Krishna, M.Stark, L.-j.Li, D.A.Shamma, M.S.Bernstein, and L.Fei-Fei. Image retrieval using scene graphs. In Computer Vision and Pattern Recognition(CVPR), 2015.

[7] J.Jhonson, A.Gupta, and L.Fei-Fei. Image generation from scene graphs. In Computer Vision and Pattern Recognition(CVPR), 2018.

[8] H.Zhang, T.Xu, H.Li, S.Zhang, X.Wang, H.Huang, and D.Metaxas. StackGAN:text to photo-realistic image synthesis with stacked generative adversarial networks. In International Conference on Computer Vision(ICCV), 2017.

[9] A.Krezhevsky, I.Sutskever, and G.E.Hinton. ImageNet classification with deep convolutional neural networks. In Neural Information Processing Systems(NIPS), 2012.

[10] I.Sutskever, O.Vinyals, and Q.V.Le. Sequence to sequence learning with neural networks. In Neural Information Processing Systems(NIPS), 2014.

[11] O.vinyals, A.Toshev, S.Bengio, and D.Erhan. Show and tell: a neural image caption generator. In Computer Vision and Pattern Recognition(CVPR), 2015.

[12] P.Anderson, X.He, C.Buehler, D.Teney, M.Johnson, S.Gould, and L.Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Computer Vision and Pattern Recognition(CVPR), 2018.

[13] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N.Gomez, L.Kaiser, and I.Polosukhin. Attention is all you need. In Neural Information Processing Systems(NIPS), 2017.

[14] Y.Feng, L.Ma, W.Liu, and J.Luo. Unsupervised image captioning. In Computer Vision and Pattern Recognition(CVPR), 2019.

[15] I.J.Goodfellow, J.P.Abadie, M.Mirza, B.Xu, D.Warde-Farley, S.Ozair, A.Courville, and Y.Bengio. Generative advesarial nets. In Neural Information Processing Systems(NIPS), 2014.

[16] D.Xu, Y.Zhu, C.B.Choy, and L.Fei-Fei. Scene graph generation by iterative message passing. In Computer Vision and Pattern Recognition(CVPR), 2017.

[17] Y.Li, W.Ouyang, B.Zhou, K.Wang, and X.Wang. Scene graph generation from objects, phrases and region captions. In International Conference on Computer Vision(ICCV), 2017.

[18] J.Yang, J.Lu, S.Lee, D.Batra, and D.Parikh. Graph r-cnn for scene graph generation. In Europian Conference on Computer Vision(ECCV), 2018.

[19] R.Zellers, M.Yatskar, S.Thomson, and Y.Choi. Neural motifs: scene graph parsing with global context. In Computer Vision and Pattern Recognition(CVPR), 2018.

[20] T.Mikolov, K.Chen, G.Corrado, and J.Dean. Efficient estimation of word representations in vector space. In International Conference on Learning Representaions(ICLR), 2013

[21] R.Krishna, Y.Zhu, O.Groth, J.Jhonson, K.Hata, J.Kravitz, and S.Chen, Y.Kalantidis, L.-j.Li, D.A.Shamma, M.S.Bernstein, and L.Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In International Journal of Computer Vision(IJCV), 2017.

[22] T.-y.Lin, M.Maire, S.Belongie, J.Hays, P.Perona, D.Ramanan, P.Dollár, and C.L.Zitnick. Microsoft coco: common objectsin context. In European Conference on Computer Vision(CVPR), 2014.

[23] G.Yin, L.Sheng, B.Liu, N.Yu, X.Wang, and J.Shao. Context and attribute grounded dense captioning. In Computer Vision and Pattern Recognition(CVPR), 2019.

[24] J.Pang, K.Chen, J.Shi, H.Feng, W.Ouyang, and D.Lin. Libra r-cnn: towards balanced learning for object detection. In Computer Vision and Pattern Recognition(CVPR), 2019.

[25] M.Pedersoli, T.Lucas, C.Schmid, and J.Verbeek. Areas of attention for image captioning. In International Conference on Computer Vision(ICCV), 2017.

[26] P.Young, A.Lai, M.Hodosh, and J.Hockenmaier. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. In Transactions of the Association for Computational Linguistics(TACL), 2014.