修士論文

注意機構による画像特徴量選択を用いた
画像付き含意関係認識

本多 右京

2019年3月15日

奈良先端科学技術大学院大学
情報科学研究科

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

本多 右京

審査委員：

松本 裕治 教授　（主指導教員）
中村 哲 教授　（副指導教員）
新保 仁 准教授　（副指導教員）
進藤 裕之 助教　（副指導教員）

注意機構による画像特徴量選択を用いた画像付き含意関係認識*

本多 右京

内容梗概

自然言語理解に視覚情報を活用する試みとして，画像を用いた含意関係認識が研究されてきた．先行研究では，言語表現はそれに紐付けられた画像の全体に対応するものとして扱われるが，言語表現の内容は必ずしも画像に写っているものすべてに対応しない．また，抽象語や機能語など，画像として表現することが難しい言語表現に画像を対応付けることは適切でない．本研究では，注意機構 (Attention) によって画像中から単語に対応した部分を選択することで，より詳細な言語表現と画像の対応付けと，その分析を行うことを目的とする．注意機構には，ダミーの特徴量を付加した，単語から画像への注意機構を用いた．ダミー特徴量付き注意機構では，通常の注意機構と同様にクエリとなる単語と対応する画像の部分を重み付けて処理できることに加えて，画像のどの部分とも対応しない単語については注意の重みをダミー特徴量へ逃がすことが可能になる．その上で，画像に対応する部分をもたない単語の情報を補完するために，画像への非対応の度合いに応じて単語特徴量を足し合わせる単語-画像特徴量ゲートを提案した．実験の結果，提案手法での画像情報の導入によって含意関係認識の精度が向上することが確認された．しかし，分析では，画像特徴量は具体的な物や行為を表す単語との対応の度合いとして限定的にしか用いられていないことがわかり，画像のより効果的な活用には課題が残ることがわかった．

キーワード

含意関係認識，注意機構，視覚と言語，記号接地

*奈良先端科学技術大学院大学 情報科学研究科 修士論文, 2019 年 3 月 15 日.

Image Feature Selection with Attention for Visually Aided Natural Language Inference*

Ukyo Honda

Abstract

Visually aided Natural Language Inference (NLI) is a challenging task to utilize the visual information for natural language understanding. In previous works, given a pair of a language expression and an image, the contents of the language expression is assumed to correspond to the whole of the image, which is not necessarily the case. Especially, function words and abstract words should not be matched to images since their contents are difficult to express in images. In this work, we aim to match words to the specific parts of images using attention, and to analyze the matches. We exploit an attention with dummy features, which allows attention not to match any parts of a image with a query word, as well as to specify the parts of the image that correspond to the query word. To compensate for the information which is difficult to be expressed in images, we propose a gate that adds word features to the selected image features, according to an incorrespondence rate between the query word and image. Our experiments show that our model processes image features effectively and outperforms the baselines which only use language inputs. Nevertheless, our in-depth analysis shows that the image features are only used as the correspondence scores with the query words in our model. We leave the more effective ways to incorporate image features in NLI as our future work.

*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, March 15, 2019.

Keywords:

natural language inference, attention, vision and language, grounding

目 次

図 目 次

表 目 次

1. はじめに

1.1 背景

人間の言語理解には，言語で表現された内容の視覚的な理解が含まれる．特に，具体的な物や出来事を指した言語表現から我々が理解する内容の多くは，実際に見えている，あるいは視覚的に思い浮かべられたそれらの物や出来事である．基本的な自然言語理解を問う含意関係認識タスクでは，画像によりこの視覚情報を補う試みがなされてきた [27, 9, 13, 7, 26, 15, 14, 31]．

これらの先行研究では，言語表現はそれに紐付けられた画像の全体に対応するものとして扱われる．しかし，言語表現の内容は必ずしも画像に写っているものすべてに対応しない．例えば，“A man is walking a dog.” というキャプションが与えられた画像には，文には言及されていない大きなビルが写っているかもしれない．この場合，このキャプションは画像の全体にではなく，画像中の男性と犬が写っている部分に対応付けられることが望ましい．また，“the” のような機能語や “ideology” のような抽象語はそもそも画像での表現が困難であり，これらの言語表現に画像を対応付けることは適切でない．先行研究では，このような言語-画像間の詳細な対応/非対応関係が考慮されていない問題点があった．また，この問題点に付随して，画像のどの部分が言語表現に対応する部分として含意関係認識に用いられているのかという，画像の使われ方の分析もなされてこなかった．

1.2 目的

本研究では，注意機構 (Attention) によって画像中から単語に対応した部分を選択することで，より詳細な言語表現と画像の対応付けと，その分析を行うことを目的とする．注意機構には，ダミーの特徴量を付加した，単語から画像への注意機構を用いた．ダミー特徴量付き注意機構では，通常の注意機構と同様にクエリとなる単語と対応する画像の部分を重み付けて処理できることに加えて，画像のどの部分とも対応しない単語については注意の重みをダミー特徴量へ逃がすことが可能になる．その上で，画像に対応する部分をもたない単語の情報を補完す

るために，画像への非対応の度合いに応じて単語特徴量を足し合わせる単語-画像特徴量ゲートを提案した．実験の結果，提案手法での画像情報の導入によって含意関係認識の精度が向上することが確認された．しかし，分析では，画像特徴量は具体的な物や行為を表す単語との対応の度合いとして限定的にしか用いられていないことがわかり，画像のより効果的な活用には課題が残ることがわかった．

1.3 本論文の構成

本論文の構成は以下のとおりである．2 章ではまず，本研究の前提として，画像を用いた含意関係認識タスクの概要を説明する．3 章では画像を用いた含意関係認識についての関連研究とその問題点を挙げる．4 章で提案手法の説明をした後，5 章で実験とその結果を示す．6 章では提案手法の挙動について分析を行う．最後に，7 章にてまとめを述べる．

2. 含意関係認識

2.1 タスク概要

本研究で取り組むタスクは含意関係認識である．含意関係認識は基本的な自然言語理解を問うタスクとして設計されており [2]，具体的なタスクは次のとおり: 前提文 (premise) と仮説文 (hypothesis) の 2 文を与えられ，前提文に対する仮説文の関係を，含意 (entailment)，矛盾 (contradiction)，無関係 (neutral) の 3 つのいずれかに分類する．例として，次の 3 対の文を考える．

前提文 (premise)
A woman with a green headscarf, blue shirt and a very big grin.

仮説文 (hypothesis)
含意 (entailment): The woman is very happy.
矛盾 (contradiction): The woman has been shot.
無関係 (neutral): The woman is young.

この例では，女性が笑っているという前提文に対して，その女性が幸せであるという仮説文は含意の関係，その女性が銃に撃たれたという仮説文は矛盾の関係，その女性が若いという仮説文は無関係と分類されるのが正しい．笑っているということは幸福感を感じている可能性が高く，逆に傷を負わされたところである可能性は低いと推論されるので，最初 2 つの仮説文はそれぞれ含意と矛盾の関係にあると分類される．3 番目の仮説文については，人間は老若男女にかかわらず笑うので，笑っている人ならば若者だろうという推論は成り立たない．しかし，笑っているということは若者ではありそうにないという推論もまた成り立たないので，3 番目の仮説文は含意でも矛盾でもなく，無関係と分類される．この例からわかるとおり，ここでの含意関係は厳密に論理的な推論に基づく関係ではなく，常識の範囲での推論に基づく関係を指す．

2.2 画像を用いた含意関係認識

含意関係認識は言語情報のみで解くことを想定されたタスクである．しかし，含意関係認識のためのデータセットはSICK[16]，SNLI[2] など画像のキャプションから構築されているものが多いことから，画像との親和性が高いタスクにもなっている．実際，これらのデータセットは他の自然言語処理タスクのデータセットと比較して具体的な事物を指す単語の割合が高いため，画像を利用することの利点が比較的大きいことが報告されている [9]．画像は，特にこれら具体的な事物に関して，言語表現とは異なる情報を与えることができる．例えば，“orange” と “grape” は同じ果物であることから似た文脈に出現し，言語表現上の比較では違いが明確にならない可能性があるが，画像表現上で比較することができれば，色や形状からこの 2 単語が指す内容 (オレンジとブドウ) の違いを明確にすることができる．また逆に，“field” と “tall green grass” のように，言語表現上は大きく異なっていても画像表現上は類似する場合などに，画像を媒介として言語表現の指す内容の類似性を考慮できることも，画像を利用することの利点として挙げられる [7].

このような背景から，[27] では，SNLI の前提文に対応する画像を加えた含意関係認識データセット，VSNLI が提案された．SNLI では，前提文がすべて画像のキャプションであるという特質から，前提文に対して元となった画像を与えることができる．ただし，仮説文は画像抜きで前提文のみを基に作成された文なので，必ずしも画像に対応するとは限らない．本研究は，この VSNLI において，画像を自然言語処理に活用する手法を検討する．

2.3 含意関係認識に画像を用いる際に考慮するべき点

2.3.1 単語レベルでの言語-画像間対応付け

含意関係認識は単語 1 つ違うだけで正解ラベルが異なりうるタスクであるため，用いる画像情報も単語レベルで文との対応をとることが望ましい．例えば “A **man** walks along with a can of soda in his hand.” という前提文と “A **woman** walks along with a can of soda in his hand.” という仮説文では，異なる部分は

“man”/“woman” の 1 単語だけである．単語が 1 つ違うだけだが，この違いによりこの 2 文は矛盾の関係と分類するのが正解とされる．しかし，文全体では 2 文はこの単語以外はすべて同じであるため，男性がソーダの缶を持って歩いている画像を与えられれば，機械はどちらの文もこの画像に対応した内容だと判断してしまう可能性が高い．この場合画像情報としては 2 文に対して同じ情報が与えられることになってしまい，矛盾であると正しく分類する妨げになってしまう．文全体ではなく各単語から画像との対応関係を考えれば，“man”/“woman” の違いを，文全体の類似性に埋もれさせることなく反映することができる．このように，1 単語の異なりが重要になる含意関係認識においては，文全体からではなく各単語から画像への対応をとることが適切であると考えられる．

2.3.2 言語-画像間の対応/非対応関係

前項で述べたとおり，含意関係認識においては単語レベルで画像との対応をとることが望ましい．しかし一方で，単語レベルでの細かい言語-画像間対応付けにおいては，単語に対応する画像の部分が存在しない可能性も考慮することが必要になる．これは，単語には抽象語や機能語など画像中に対応する部分を見つけることが難しいものが存在することによる．単語レベルで細かく画像との対応をとる場合には，画像のどの部分にも対応しない単語に誤った画像情報を与えないよう配慮しなければならない．また，特に VSNLI というデータセットの性質であるが，文自体が画像と対応していない場合も存在するため，この場合に無理にその画像の一部分を単語に対応する箇所として与えないよう工夫する必要がある．VSNLI では矛盾や無関係の関係にある仮説文は画像と異なる内容を述べている可能性が高く，例えば画像には男性しかいないのに “There is a woman.” という仮説文が与えられた場合，この仮説文の内容は画像とは異なっており “woman” という単語に対応する画像の部分は存在しない．この場合，画像のどの部分を “woman” という単語に与えても誤りとなってしまうため，この単語は画像のどの部分にも対応しない，と判断できなければならない．加えて，言語表現が画像の全体には対応せず部分的にしか対応しない場合も考えられる．例えば，“dog” という単語に対して犬と人間が写った画像が与えられた場合，“dog” は画像中の犬の写った

部分だけを指すのだということがわからなければ，人間を写した部分の画像特徴量がノイズとして後続の処理に渡されてしまう．この場合も，“dog” は犬の写った部分と対応するのであり，人間の写った部分とは対応しないという判断が必要となる．このように，画像を用いた含意関係認識においては，単語-画像間の対応関係と同時に，非対応関係の判断も求められる．

3. 関連研究

画像を用いた含意関係認識では，画像の使い方について次のような手法が提案されてきた．[31, 14, 15] では，語句に対応した画像はキャプション中にその語句を含む画像のセットとされる．語句の含意関係は，語句のもつ画像のセット間で画像の一致数や異なり数を見ることで，同じような状況 (画像) を表した語句か否かを基に分類される．[9] では，キャプション全体とその元の画像全体が対応しているものとして，キャプションをエンコードして得た特徴量で元画像を識別できるよう，エンコーダを学習させる．これにより事前学習でエンコーダに画像特徴量の分布を反映させ，このエンコーダの出力を転移学習させて含意関係認識タスクを解く．[7, 13] は，これまでのキャプションデータセットを使った研究とは異なり，語句による画像検索の結果をその語句に対応する画像のセットとして利用する手法を提案している．これにより，キャプションデータセットの語彙や画像に限定されることなく，言語表現に対応した画像を得ることができる．

上記のいずれの研究も，言語表現は画像全体に対応するものとして扱っている．しかし，2.3.2 節で述べたとおり，言語表現には画像と対応関係をもたないものや，部分的にしか対応しないものが考えられる．部分的にこれを考慮したものとして，[26] と [27] が挙げられる．[26] では，言語表現はそれに対応する画像に対して上位の階層関係にあるとして，言語表現と画像全体の対応を学習させるのではなく，言語表現が画像の上位に階層にあるという順序関係を学習させる．[27] は，前提文と仮説文の対に，キャプションである前提文の元となった画像を加えたデータセットにおいて，前提文-仮説文間，仮説文-画像の各区画間でコサイン類似度をとり，この類似度に基づいて含意関係を判断する手法を提案した．これらは言語表現を画像の全体と対応させているわけではない点で言語-画像間の非対応の関係も考慮しているといえるが，[26] は順序関係，[27] はコサイン類似度と，画像の特徴量を限定的にしか用いない．本研究では注意機構によって選択した画像の部分の特徴量を使うため，より直接的に画像特徴量を使うことが可能なモデルとなっており，また注意の重みの分析によって画像の使われ方の解釈を行う可能性をもつ点で，これらの先行研究と異なる．

4. 提案手法

本章では，提案手法を，基本特徴量，注意機構，ゲート，文表現の 4 つの段階に分けて説明していく．基本特徴量，注意機構，ゲートで文中の各単語の特徴量を計算し，文表現ではそれを文の特徴量としてまとめる．なお，本章末の図 1 に提案手法の概要を示してある．

4.1 基本特徴量

2 章で述べたとおり，本研究で扱うデータは VSNLI[27] である．VSNLI のデータセットが与える情報は，1 設問ごとに，{ 前提文，仮説文，前提文と仮説文の関係の正解ラベル，前提文に対応する画像 } の 4 つである．文，画像は特徴量に変換し，I 個の単語からなる文の特徴量を $S = [\mathbf{s}_1, ..., \mathbf{s}_I] \in \mathbb{R}^{d_{word} \times I}$，$J$ 個の区画からなる画像の特徴量を $V = [\mathbf{v}_1, ..., \mathbf{v}_J] \in \mathbb{R}^{d_{img} \times J}$ とする．後述する注意機構におけるクエリには，この単語特徴量 $\mathbf{s}_i$ を用いる．各特徴量のエンコーダの詳細は，5 章の実験設定に示す．

単語と画像の特徴量は事前学習された値を固定し，学習可能なパラメータ $W_s^{base} \in \mathbb{R}^{d \times d_{word}}$, $W_v^{base} \in \mathbb{R}^{d \times d_{img}}$, $\mathbf{b}_s^{base} \in \mathbb{R}^d$, $\mathbf{b}_v^{base} \in \mathbb{R}^d$ により次のように変換する：

$$\mathbf{s}_i^{base} = \tanh(W_s^{base}\mathbf{s}_i + \mathbf{b}_s^{base}), \quad (1)$$

$$\mathbf{v}_j^{base} = \tanh(W_v^{base}\mathbf{v}_j + \mathbf{b}_v^{base}). \quad (2)$$

この $\mathbf{s}_i^{base}$, $\mathbf{v}_j^{base}$ を単語と画像の基本的な特徴量とする．

4.2 注意機構

注意機構 (Attention)[17, 1] は，入力の特徴量のうちクエリとなる特徴量と関連の強いものに重みを付けて処理する手法である．言語と画像の融合領域においても，画像の入力のうちクエリの言語表現に関連の強い入力に重みを付けて処理する用途で使用されており，キャプション生成や Visual Question Answering (VQA)

で有効であることが報告されている [30, 22]．提案手法でも，この注意機構を使って単語に関連の強い画像の特徴量を選択する．

単語をクエリとして画像中でこれに関連する部分を特定する注意機構は，単語-画像間の対応関係を判断する役割を担っていると考えられる．2.3.2 項で述べたように，単語-画像間の対応を考える際は，対応する画像の部分の判断と同時に，画像が対応しないことの判断も必要になる．提案手法では，ダミー特徴量付きの注意機構を用いることによって，単語-画像間の対応，非対応関係の両側面を考慮する．以下で具体的にこれを説明する．

4.2.1 基本的な注意機構

基本的な注意機構での計算は次のように行われる．文中の i 番目の単語をクエリとしてその特徴量を $\mathbf{s}_i$ とし，入力画像の j 番目の区画の特徴量を $\mathbf{v}_j$ とすると，i 番目の単語から入力画像の j 番目の区画への注意の重み $\alpha_{i,j}$ は，

$$\alpha_{i,j} = \frac{\exp(\langle \mathbf{s}_i, \mathbf{v}_j \rangle)}{\sum_{k=1}^{J} \exp(\langle \mathbf{s}_i, \mathbf{v}_k \rangle)}. \tag{3}$$

$\langle \cdot, \cdot \rangle$ は内積を表す．上式のとおり，注意機構の重みはクエリと入力の特徴量の内積に softmax 関数を適用した値になっており，内積が大きい入力ほど重みの値が大きくなる．注意機構は，この重みの大きさをクエリと入力との対応の度合いと見做すことで，クエリに対応する入力を大きく重み付けして処理する．注意機構の出力は，入力に式 (3) で得られる重みをかけた値となっており，i 番目の単語から入力画像への注意機構の出力を $\hat{\mathbf{v}}_i$ とすると，

$$\hat{\mathbf{v}}_i = \sum_{j=1}^{J} \alpha_{i,j} \mathbf{v}_j. \tag{4}$$

4.2.2 複数注意機構

以上が注意機構での基本的な計算であるが，最近の研究では，注意機構の重みはクエリ，入力を変換することにより複数回求めることが有効であることが報告されている [25]．この複数注意機構は言語から画像への注意においても有効であ

ることが報告がされている [10, 18, 11]．これを踏まえて，本研究では次のような複数注意機構を用いる．

M 個の注意機構があるとして，このうち m 番目の注意機構における，文中 i 番目の単語から j 番目の画像区画への注意の重みを $\alpha_{i,j}^{(m)}$ とすると，$\alpha_{i,j}^{(m)}$ は次のように計算される：

$$\mathbf{s}_i^{att} = \tanh(W_s^{att}\mathbf{s}_i^{base} + \mathbf{b}_s^{att}), \tag{5}$$

$$\mathbf{v}_j^{att} = \tanh(W_v^{att}\mathbf{v}_j^{base} + \mathbf{b}_v^{att}), \tag{6}$$

$$e_{i,j}^{(m)} = \exp(\frac{\langle \mathbf{q}^{(m)}, \mathbf{s}_i^{att} \odot \mathbf{v}_j^{att} \rangle}{\sqrt{d}}), \tag{7}$$

$$\alpha_{i,j}^{(m)} = \frac{e_{i,j}^{(m)}}{\sum_{k=1}^{J} e_{i,k}^{(m)}}. \tag{8}$$

ここで，$W_s^{att} \in \mathbb{R}^{d\times d}$, $W_v^{att} \in \mathbb{R}^{d\times d}$, $\mathbf{b}_s^{att} \in \mathbb{R}^d$, $\mathbf{b}_v^{att} \in \mathbb{R}^d$, $\mathbf{q}^{(m)} \in \mathbb{R}^d$ はすべて学習可能なパラメータである．$\odot$ は要素積を表す．d は画像入力の特徴量の次元数である．上記の変換により，同じクエリと入力から，M 個の異なる注意の重みを得ることが可能になる．最終的な注意の重み $\alpha_{i,j}$ は M 個の注意の重みの平均をとって，

$$\alpha_{i,j} = \frac{1}{M}\sum_{m=1}^{M} \alpha_{i,j}^{(m)}. \tag{9}$$

i 番目の単語から入力画像への注意機構の出力 $\hat{\mathbf{v}}_i$ は式 (4) と同様に，

$$\hat{\mathbf{v}}_i = \sum_{j=1}^{J} \alpha_{i,j}\mathbf{v}_j^{base}. \tag{10}$$

4.2.3 ダミー特徴量付き注意機構

前述のとおり，注意機構は単語-画像間の対応関係を判断する役割を担っていると考えられる．しかし，これまでの注意機構では 2.3.2 項で述べた，単語-画像間の非対応の判断が行えない問題がある．注意の重みはクエリと画像の特徴量の内積に softmax 関数を適用したものであり，どんなクエリであっても画像への注意

の重みの和は必ず 1 になる．すなわち，この注意機構ではどんなクエリであっても必ず，画像のいずれかの部分を対応する部分として重み付けすることになる．この注意機構の性質が，画像に対応部分をもたない単語にも無理に誤った画像情報を与えてしまう問題を引き起こす．単語の中には “ideology” のような抽象語や “the” などの機能語といった，画像として表現することが困難な単語が存在する．また，特に VSNLI というデータセットの性質であるが，文自体が画像と対応していない場合も存在し，この場合文中の単語が画像に対応しないことが考えられる．これらの場合にも，上記の注意機構では必ず画像のいずれかの部分を単語に対応する画像情報として与えてしまうため，ノイズの多い画像特徴量選択になってしまう可能性がある．

この問題を解決するために，本研究ではダミーの特徴量を付加した注意機構を用いる．これは元々質問応答タスクで導入されたもので [29]，画像と言語にまたがる質問応答である VQA においても，互いに対応のない単語/画像に対して注意の重みをかけないための工夫として用いられている [18]．この手法では，画像特徴量にダミーの特徴量を加えることで，画像に対応部分をもたない単語がこのダミーの特徴量に注意の重みをおくことを可能にする．ダミーの特徴量は，K 個の学習可能なパラメータ $P = [\mathbf{p}_1, ..., \mathbf{p}_K] \in \mathbb{R}^{d_{img} \times K}$ とする．この P を画像特徴量 V に加えることによって，ダミーの特徴量を加えた画像特徴量 $V^{dummy} = [\mathbf{v}_1, ..., \mathbf{v}_J, \mathbf{p}_1, ..., \mathbf{p}_K] \in \mathbb{R}^{d \times (J+K)}$ が得られる．この V^{dummy} を式 (2) と同じパラメータで非線形変換したものを $V^{base'}$ とすると，

$$V^{base'} = \tanh(W_v^{base} V^{dummy} + \mathbf{b}_v^{base}). \tag{11}$$

ダミー特徴量付き注意機構では，この $V^{base'}$ を用いて，4.2.2 項の注意の重みの計

算を以下のように変更する：

$$\mathbf{v}_l^{att'} = \tanh(W_v^{att}\mathbf{v}_l^{base'} + \mathbf{b}_v^{att}), \tag{12}$$

$$e_{i,l}^{(m)} = \exp(\frac{\langle \mathbf{q}^{(m)}, \mathbf{s}_i^{att} \odot \mathbf{v}_l^{att'} \rangle}{\sqrt{d}}), \tag{13}$$

$$\alpha_{i,l}^{(m)} = \frac{e_{i,l}^{(m)}}{\sum_{n=1}^{J+K} e_{i,n}^{(m)}}, \tag{14}$$

$$\alpha_{i,l} = \frac{1}{M}\sum_{m=1}^{M} \alpha_{i,l}^{(m)}. \tag{15}$$

ここまで，注意の重みは，画像特徴量 V にダミーの特徴量 P を加えた $J+K$ 個の画像特徴量について計算してきた．式 (15) は softmax の出力であるから，画像との内積値が小さければ，相対的にダミーへの注意の重みが大きくなる．このダミーへの注意の重みを注意機構の出力の計算時には除外することによって，画像と対応がない単語をクエリとした場合には画像への注意の重みの総和を小さくすることが可能になる．i 番目の単語から入力画像への注意機構の出力 $\hat{\mathbf{v}}_i$ は，

$$\hat{\mathbf{v}}_i = \sum_{j=1}^{J} \alpha_{i,j}\mathbf{v}_j^{base}. \tag{16}$$

式 (16) ではダミーの特徴量も，ダミーの特徴量にかかった重みも用いられない $(\mathbf{v}_j^{base};\ \alpha_{i,j};\ j = 1, ..., J)$．このため，ダミーへの注意の重みが大きければ大きいほど画像特徴量への注意の重みの総和が小さくなり，注意機構の出力に加算する画像特徴量が少なくなることになる．

4.3 ゲート

前節では，ダミー特徴量付き注意機構により，画像との非対応の可能性も考慮した，単語による画像特徴量の選択を行った．これにより，画像で表現することが難しい単語にも無理に画像特徴量を与えてノイズの多い特徴量を割り当ててしまう問題は軽減されると考えられる．しかし，この選択の結果だけでは，画像中に対応する部分をもつことが難しい抽象語や機能語については，単に特徴量を与えず無視することになってしまうおそれがある．抽象語や機能語は，画像で表現

することが難しいだけで，文の内容においては重要な役割を果たす．例えば，2.1節で挙げた含意関係認識の例で，仮説文の “happy” という抽象語が無視されてしまうと，女の人が笑っているという前提文との含意の関係を正しく分類することができなくなってしまう．

そこで本研究では，画像で表現することが難しい単語についてはその難しさの度合いに応じて単語特徴量を付け加える，単語-画像特徴量ゲートを提案する．この難しさの指標には，前節の注意機構では直接使われなかったダミーの特徴量への注意の重みを用いる．ダミーの特徴量への注意の重みは，大きければ大きいほど単語が画像に対応する部分をもたず，画像で表現することが困難であることを示す指標とみなせる．この指標を用いて，式 (1) の単語特徴量と，式 (16) の注意機構出力の画像特徴量との次のような重み付き和 $\mathbf{g}_i$ をとる：

$$\gamma_i = \sum_{o=J+1}^{J+K} \alpha_{i,o}, \tag{17}$$

$$\phi_s(\mathbf{x}) = \mathrm{ReLU}(W_s^{gate}\mathbf{x} + \mathbf{b}_s^{gate}), \tag{18}$$

$$\phi_v(\mathbf{x}) = \mathrm{ReLU}(W_v^{gate}\mathbf{x} + \mathbf{b}_v^{gate}), \tag{19}$$

$$\mathbf{g}_i = \gamma_i \phi_s(\mathbf{s}_i^{base}) + (1-\gamma_i)\phi_v(\hat{\mathbf{v}}_i). \tag{20}$$

$W_s^{gate} \in \mathbb{R}^{d\times d}$, $W_v^{gate} \in \mathbb{R}^{d\times d}$, $\mathbf{b}_s^{gate} \in \mathbb{R}^d$, $\mathbf{b}_v^{gate} \in \mathbb{R}^d$ はすべて学習可能なパラメータである．上記の計算で，画像で表現可能と判断された単語については画像特徴量を多く，言語でしか表現できないと判断された単語については単語特徴量を多く加えた単語特徴量が得られる．このように，単語-画像特徴量ゲートでは，ダミー特徴量付き注意機構で犠牲にされる，言語でしか表現できない情報を補完する役割が期待される．

4.4 文表現

前節までで，画像特徴量と重み付き和をとった単語特徴量 $\mathbf{g}_i$ が得られる．提案手法の中心部分は，ここまでの，ダミー特徴量付き注意機構による画像特徴量の選択と，単語-画像特徴量ゲートによる単語特徴量の補完である．これ以降の処理は，言語表現のみを用いた含意関係認識モデルである InferSent[4] において，入

力の単語特徴量 $\mathbf{s}_i$ を $\mathbf{g}_i$ に置き換えた処理に相当する．以下でこれを具体的に説明する．

文の各単語について，式 (20) の重み付き和を計算し，$[\mathbf{g}_1, ..., \mathbf{g}_I] \in \mathbb{R}^{d \times I}$ を得たとする．これを文頭からと文末からの双方向の LSTM[8](BiLSTM) に通して，文の語順情報をエンコードする．語順情報をエンコードされた各単語の特徴量を $\mathbf{h}_i$ とし，これは BiLSTM の各方向の隠れ層を次のように連結したものとする：

$$\mathbf{h}_i^{forward} = \mathrm{LSTM}^{forward}(\mathbf{g}_i, \mathbf{h}_{i-1}^{forward}), \tag{21}$$

$$\mathbf{h}_i^{backward} = \mathrm{LSTM}^{backward}(\mathbf{g}_i, \mathbf{h}_{i+1}^{backward}), \tag{22}$$

$$\mathbf{h}_i = [\mathbf{h}_i^{forward}; \mathbf{h}_i^{backward}]. \tag{23}$$

$[\cdot;\cdot]$ は連結 (concatenate) を表す．各特徴量の次元数については，$\mathbf{h}_i^{forward} \in \mathbb{R}^{d/2}$，$\mathbf{h}_i^{backward} \in \mathbb{R}^{d/2}$，$\mathbf{h}_i \in \mathbb{R}^d$．$\{\mathbf{h}_i\}_{i=1}^I$ の各次元の平均値と最大値をとったベクトルの連結を，文全体を表す特徴量 $\mathbf{r}$ とすると，

$$\mathbf{r} = [\mathrm{mean}(\{\mathbf{h}_i\}_{i=1}^I); \mathrm{max}(\{\mathbf{h}_i\}_{i=1}^I)]. \tag{24}$$

$\mathrm{mean}(\cdot)$，$\mathrm{max}(\cdot)$ はそれぞれベクトルの集合から各次元における平均値と最大値を返す関数とする．$\mathbf{r} \in \mathbb{R}^{2d}$．前提文と仮説文それぞれについて式 (24) を適用し，それぞれの文の特徴量 $\mathbf{r}_{pre}$，$\mathbf{r}_{hyp}$ を得る．2 文の関係を表す特徴量を $\mathbf{z}$ とし，次式で計算する：

$$\mathbf{z} = [\mathbf{r}_{pre}; \mathbf{r}_{hyp}; \mathbf{r}_{pre} \odot \mathbf{r}_{hyp}; |\mathbf{r}_{pre} - \mathbf{r}_{hyp}|]. \tag{25}$$

$\mathbf{z} \in \mathbb{R}^{8d}$．次に，$\mathbf{z}$ を多層パーセプトロン (MLP) に通して 3 次元のベクトル $\mathbf{u} \in \mathbb{R}^3$ に変換し，これに softmax 関数を適用する．これにより得られたベクトル $\mathbf{a} \in \mathbb{R}^3$ の各次元の値を，それぞれ 2 文の関係が含意，矛盾，無関係であると推定した確率とみなす．$\mathbf{u}$，$\mathbf{a}$ の l 番目の要素をそれぞれ u_l，a_l とすると，

$$\mathbf{u} = \mathrm{MLP}(\mathbf{z}), \tag{26}$$

$$a_l = \frac{\exp(u_l)}{\sum_{m=1}^3 \exp(u_m)}. \tag{27}$$

学習は，推定された確率分布 $\mathbf{a}$ と真の確率分布との交差エントロピー誤差 E を最小化するよう行う．正解となる関係についての確率が $\mathbf{a}$ の c 番目の次元の値 a_c であるとすると，

$$E = -\log a_c. \tag{28}$$

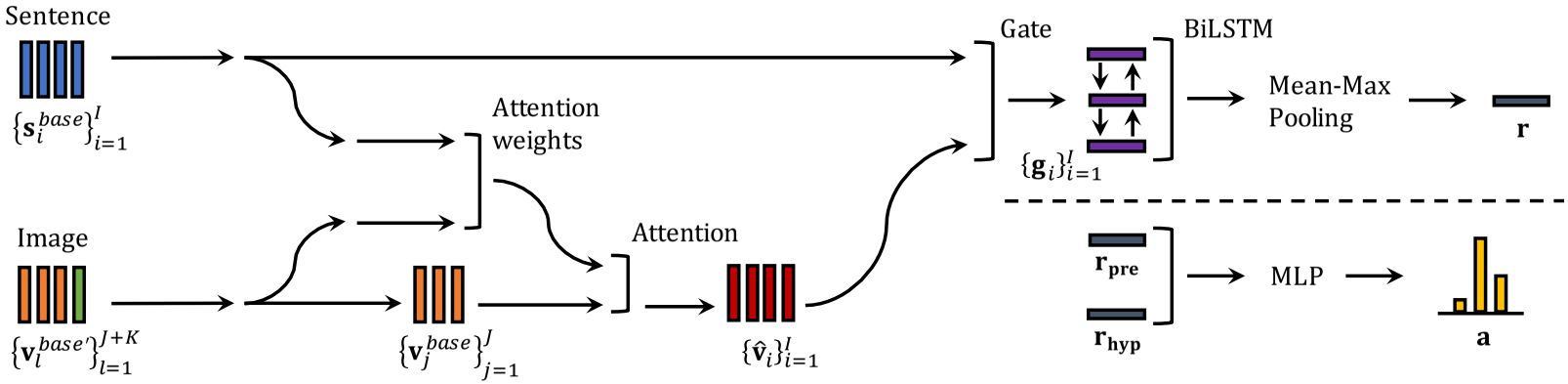

Sentence
$\{\mathbf{s}_i^{base}\}_{i=1}^{I}$
Image
$\{\mathbf{v}_l^{base}\}_{l=1}^{J+K}$
Attention weights
$\{\mathbf{v}_j^{base}\}_{j=1}^{J}$
Attention
$\{\hat{\mathbf{v}}_i\}_{i=1}^{I}$
Gate
$\{\mathbf{g}_i\}_{i=1}^{I}$
BiLSTM
Mean-Max Pooling
$\mathbf{r}$
$\mathbf{r_{pre}}$
$\mathbf{r_{hyp}}$
MLP
$\mathbf{a}$

図 1: 提案手法概要

5. 実験

5.1 データセット

データセットはVSNLI[27] を使った．データ数，ラベル数の統計は表 1 のとおりである．表 1 から，データセット区分 (train, dev, test, test$_{hard}$) ごとに各ラベル (含意，矛盾，無関係) に属する設問の数が等しくなっていることがわかる．このため，VSNLI での評価には精度 (accuracy) が用いられる．VSNLI は SNLI[2] の一部に画像データを加えたものであるが，[6] では SNLI にはデータの偏りがあることが報告されており，同論文でこの偏りを取り除いた SNLI test$_{hard}$ のデータセットが公開されている．VSNLI test$_{hard}$ は，この SNLI test$_{hard}$ の一部に画像を付加したものである．

	含意	矛盾	無関係	合計
train	182,167	181,938	181,515	545,620
dev	3,329	3,278	3,235	9,842
test	3,368	3,237	3,219	9,824
test$_{hard}$	1,058	1,135	1,068	3,261

表 1: VSNLI データセット統計

5.2 提案手法

特徴量エンコーダ：単語ベクトルには事前学習済みの Glove[19](300 次元)，画像エンコーダーには ImageNet[5] で事前学習済みの VGG16[23] の最終 pooling 層の出力 (channel, height, width = 512, 7, 7) を用いた．どちらもパラメータを固定し，学習中に更新を行わない．Glove の語彙数は 30 万とした．

注意機構：注意機構の数 M とダミー特徴量の数 K は，先行研究 [18] にしたがってそれぞれ $M = 4$，$K = 3$ とした．

中間層：基本特徴量の次元数 d は 512，BiLSTM 隠れ層の次元数は各方向 256 次元で，合計 512 次元とした．含意関係ラベルの予測確率を出力する式 (26) の MLP は 3 層で，中間の次元数は 512，活性化関数には Swish[20] を用いた．この MLP と BiLSTM を除いた，他のすべての非線形変換において，比率 0.2 の dropout[24] を行った．
Optimizer:Optimizer には AMSGrad[21] を使用し，ハイパーパラメータは Adam[12] の推奨値に設定した．

5.3 比較手法

比較手法として，まず VSNLI の論文でのベースライン [27] を説明する．ここでは概要のみ述べるので，詳細は [27] を参照されたい．なお，ここで用いられる LSTM の隠れ層の次元数は，提案手法の BiLSTM の両方向次元数の合計と同じ，512 である．単語特徴量/画像特徴量エンコーダについても提案手法と同じ，事前学習済みの Glove(語彙数 30 万，300 次元)/VGG16 を用いている．
LSTM$_{hyp}$：前節で述べたとおり，SNLI のデータセットにはデータの偏りがあり，仮説文だけを使っても高い精度が出てしまうことが報告されている [6]．LSTM$_{hyp}$ では，仮説文を LSTM で 512 次元のベクトルにエンコードし，これを MLP に通して 3 値分類する．
LSTM：この手法では，前提文と仮説文をそれぞれ LSTM でエンコードし，この 2 つのベクトルを連結して MLP に通す．
V-LSTM：この手法では，前述の LSTM の MLP の直前に，画像の VGG16 の全結合層の出力 (4096 次元) を連結する処理を行う．
VQA：この手法では，まず事前学習した物体検出手法で画像から 36 個の物体を検出する．検出した各物体に対応した画像特徴量 (2048 次元) に対して，LSTM で文をエンコードして得たベクトルをクエリとして注意機構を適用する．クエリの特徴量と注意機構の出力で要素積をとって，文の特徴量とする．以上の文特徴量を前提文と仮説文のそれぞれに対して計算し，これらの連結を MLP に通す．
BiMPM[28]：この手法では．BiLSTM でエンコードした前提文と仮説文の間で相互にコサイン類似度をとり，このコサイン類似度をさらに BiLSTM でエンコー

ドし，最終タイムステップの隠れ層を MLP に通す．このコサイン類似度の計算の仕方を変えたモデルを複数用意し，これらのモデルのアンサンブルで予測を行う．提案手法ではこれに対応するアンサンブルを行えないので，提案手法をアンサンブルした上での比較は行っていない．

V-BiMPM：この手法では，BiMPM での前提文-仮説文間でのコサイン類似度に加えて，仮説文-画像間でのコサイン類似度も用いる．画像特徴量は提案手法と同じ，VGG16 の最終 pooling 層の出力 (channel, height, width = 512, 7, 7) を用いる．BiMPM と同様に，この手法に対しても，提案手法のアンサンブルの実装，比較は行っていない．

これらのベースラインのうち，V-LSTM, VQA, V-BiMPM は画像を使う手法，それ以外は文のみを入力とする手法である．BiMPM と V-BiMPM はアンサンブルモデル，それ以外は単体モデルである．

以上のベースラインに加えて，提案手法では次の 3 つのベースラインと比較を行った．なお，提案手法と以下の比較手法のテストデータでのスコアは，いずれも 20 epoch の学習のうち dev で最も高い精度を出したモデルについて計算した．特徴量エンコーダ，注意機構，中間層，optimizer の設定は，すべて提案手法のものとそろえた．

InferSent: この手法では，式 (21) からの計算において $\mathbf{g}$ ではなく単語の特徴量 $\mathbf{s}$ を用いる．これは通常の InferSent[4] と同じほぼ処理である．提案手法を文のみを用いるよう改変したベースラインとなっている．

InferSent + VGG16 FC: この手法では，上記 InferSent の MLP の直前に，画像の VGG16 の全結合層の出力 (4096 次元) を連結する処理を行う．注意機構による画像特徴量選択を行わなず画像を用いるベースラインとなっている．

SentenceAttention: この手法では，単語でなく文から画像への注意機構を用いる．上記 InferSent のベースラインで得た文の特徴量をクエリとして 4.2.3 項のダミー特徴量付き注意機構での処理を行い，得られた注意機構の出力と文の特徴量を 4.3 節の単語-画像特徴量ゲートに通す．以降の処理は提案手法と同様である．注意機構による画像特徴量選択を，単語レベルでなく文レベルで行うベースラインとなっている．

5.4 結果

実験結果を表 2 に示す．まず見てとれることとして，提案手法以外の単体モデルでは，画像特徴量を加えると言語の特徴量のみを使う手法より精度が下がってしまっていることが挙げられる．このことから，画像特徴量は導入を工夫しないと単にノイズになってしまうことがわかる．SentenceAttention は提案手法の単語レベルの注意機構を文レベルの注意機構にしただけの手法であるが，精度は提案手法から大きく下がっており，2.3.1 項で述べた単語レベルでの画像対応付けが有効であることが示唆される．提案手法は，test では言語の特徴量のみを使う InferSent に劣るが，データの偏りの少ない test_{hard} では InferSent より約 1 ポイント精度が高い．この難易度の高いテストセットで精度の向上は，提案手法による画像特徴量の導入の有効性を示していると考えられる．アンサンブルモデルには精度で劣る結果となったが，これについては次章でより詳細な比較を示す．

		test	test_{hard}
Single	LSTM^{*}_{hyp}	68.49	25.57
	LSTM*	81.49	60.99
	V-LSTM*	75.70	49.03
	VQA*	79.65	56.21
	InferSent	**84.78**	69.24
	InferSent + VGG16 FC	83.70	66.51
	SentenceAttention	79.50	60.88
	Proposed	84.76	**70.18**
Ensemble	BiMPM*	86.41	72.55
	V-BiMPM*	**86.99**	**73.75**

表 2: VSNLI における精度比較．* は [27] の数値．それ以外は 5 回の試行の平均値．

6. 分析

6.1 必要知識項目ごとの精度比較

	Tag	Ins	Gen	Ent	Verb	WK	Quant
Single	InferSent	62	**91**	74	74	68	**77**
	Proposed	**64**	90	**78**	**76**	**69**	**77**
Ensemble	BiMPM*	58	**93**	**95**	**77**	**79**	**78**
	V-BiMPM*	**63**	89	78	68	71	70

表 3: 必要知識項目ごとの精度比較 (test_{hard})．Tag は必要知識項目名またはその略称で，それぞれ，Ins:Insertion, Gen:Generalisation, Ent:Entity, Verb:Verb, WK:World Knowledge, Quant:Quantifier である．* は [27] の数値．それ以外は 5 回の試行の平均値．

VSNLI では，テストデータの一部に，その設問を解くために必要な知識が項目別に人手でアノテーションされている [27]．この項目別に，画像を用いた提案手法が，どのような知識を要求する設問について数値を上げているのかを検証する．項目は，[27] で BiMPM，V-BiMPM の数値が示されているもののうち，10 以上の設問数をもつものを用いた．これらの項目は，含意関係の分類にはそれぞれ次の知識が必要なことを示す : Insertion では仮説文に前提文にない余計な要素が足されていること，Generalisation では前提文が仮説文の一般化した表現であること，Entity では 2 文間で具体的な物が異なること，Verb では 2 文間で行われている行為が異なること，World Knowledge では常識とされるような世界知識，Quantifier では 2 文間で数詞が異なることが，含意関係の分類に必要とされる．

表 3 に，test_{hard} における項目別の精度を示す．この表から，文のみを入力とする InferSent のベースラインの対して提案手法の精度が上がっている項目は，Insertion，Entity，Verb，World Knowledge の知識を必要とする設問であることがわかる．このうち，Insertion, Entity, Verb については，提案手法が V-BiMPM より高い精度を示している．特に Entity と Verb で要求される具体的な物や行為

についての知識は，画像で最も表現されやすい情報であると考えられ，この項目で高い精度を示していることから，提案手法が適切に画像特徴量を考慮していることが示唆される．

6.2 要素ごとの効果の比較

	test	test_{hard}
Full	**85.00**	**70.75**
-Gate	83.12	67.99
-Gate&-Dummy	82.94	67.40

表 4: 要素ごとの効果の比較

提案手法で中心的な役割を果たすのは，ダミー特徴量付き注意機構と単語-画像特徴量ゲートである．ここではこの 2 つの要素それぞれの，精度に対する効果を分析する．結果を表 4 に示す．表 4 中の Full は提案手法，-Gate は，式 (20) で γ_i の重みをかけず単純な和をとったもの，-Gate&-Dummy は，上記の変更に加えてダミーの特徴量 $K = 0$ としたものである．なお，この実験では一回の試行の結果のみで比較しているため，Full のモデルの精度は表 2 の提案手法のものと異なる．

表 4 では-Gate の設定で Full から大きく精度が下がっており，特に単語-画像特徴量ゲートが精度に寄与していることがわかる．-Gate&-Dummy では-Gate よりやや数値が下がるに留まっており，ダミー特徴量付き注意機構による画像特徴量の操作だけではそれほど機能していないことがわかる．

6.3 画像入れ替えによる比較

ここでは提案手法における画像の効果をより詳細に検証するため，前提文と仮説文のペアに対して本来与えられる画像を別の画像に入れ替えて，提案手法の精度が下がるか否かの検証を行う．入れ替える先の画像は，dev, test, test_{hard} には

	test	test_{hard}
Original	**84.76**	**70.18**
Foil	84.69	70.07

表 5: 画像入れ替え時の VSNLI における精度比較．数値は 5 回の試行の平均値．

Tag	Ins	Gen	Ent	Verb	WK	Quant
Original	64	**90**	**78**	**76**	**69**	**77**
Foil	**65**	**90**	77	74	**69**	76

表 6: 画像入れ替え時の必要知識項目ごとの精度比較 (test_{hard})．Tag は必要知識項目名またはその略称で，それぞれ，Ins:Insertion, Gen:Generalisation, Ent:Entity, Verb:Verb, WK:World Knowledge, Quant:Quantifier である．数値は 5 回の試行の平均値．

なく train でしか出現しない画像を一枚ランダムに選択した．dev, test, test_{hard} で画像をすべてこの画像に入れ替えたテストセットを作成し，学習済みの提案モデルで，パラメータを更新せずにこのテストセットでの精度を算出した．提案手法が画像の情報を活用していれば，この画像の入れ替えによって提案手法の精度が下がることが期待される．なお，以下，入れ替えた画像は [27] に倣って “foil image” と表記することがある．

表 5 に結果を示す．表 5 中の Original は元々与えられていた画像を使った際の提案手法の精度，Foil は入れ替えた画像を使った際の提案手法の精度である．この表からわかるように，若干精度の減少が見られるものの，期待に反して提案手法では画像の入れ替えによってほぼ精度が下がらない．このことから，提案手法では画像の情報を僅かにしか活用していないことがわかる．

次に，より詳細に画像入れ替えによる効果を検証するため，6.1 節でみた必要知識項目ごとに精度の比較を行った．表 6 に結果を示す．表 5 と同様，表 6 中の Original は本来与えられる画像を使った際の提案手法の精度，Foil は入れ替えた画像を使った際の提案手法の精度である．

表 6 から，全体的に画像の入れ替えによってはほぼ精度が下がらない傾向が見られるが，Entity と Verb の項目については若干精度が下がっていることがわかる．前述のとおり，Entity と Verb で要求される具体的な物や行為についての知識は，最も画像で表現されやすい情報であると考えられる．提案手法は，若干ではあるが，画像から具体的な物や行為に関する情報を得ていることが示唆される．

6.4 ゲートにおける単語-画像特徴量の比較

前節の結果は，提案手法では画像の情報を僅かにしか活用できておらず，かろうじて具体的な物や行為に関する情報を得ていることを示唆していた．本節では，なぜ画像の入れ替えによって精度がほぼ変わらないのかを，画像特徴量の値に注目して検証する．画像の入れ替えがほぼ精度に影響しないことから，提案手法のどこかで画像特徴量がほぼ 0 になっていることが考えられる．ここではゲート直前の画像特徴量について，L2 ノルムの大きさを測定する．

学習済みの提案モデルにおいて，式 (20) でゲートの重み付き和をとる前の $\phi_s(\mathbf{s}_i^{base})$，$\phi_v(\hat{\mathbf{v}}_i)$ について，dev と test に出てくる全単語の L2 ノルムをとり，その平均を算出した[1]．$\phi_s(\mathbf{s}_i^{base})$ の L2 ノルムの平均が 40.0470 であるのに対して，$\phi_v(\hat{\mathbf{v}}_i)$ の L2 ノルムの平均は 0.0380 であった．この L2 ノルムについて見れば，単語特徴量に対する画像特徴量の比は 0.0009 でほぼ 0 となり，画像特徴量はこれ以降の処理にほとんど影響しないことがわかる．よって，式 (20) での計算は，学習済み提案モデル上では実質，

$$\mathbf{g}_i \approx \gamma_i \phi_s(\mathbf{s}_i^{base}). \tag{29}$$

これにより，InferSent のベースラインに対する提案手法の精度の向上は，ダミーの特徴量への注意の重みによる，単語特徴量への重み付けであることがわかる．

これと前節での考察から，提案手法では，画像特徴量自体は分類に用いないが，画像中で具体的な物や行為に該当する部分の特徴量を使ってダミーの特徴量への重みを操作することで，画像との関係を利用していることが推察される．そこで，次節ではダミーの特徴量への重みを分析する．

[1] test_{hard} は test から設問を選別したものなので，繰り返しを避けるため除外．

6.5 ダミーへの注意の重みの分析

ここでは，単語の表す概念の具体性を人手でアノテーションした concreteness スコア [3] を使って，どのような単語がダミーの特徴量へ注意の重みをかけるかを分析した．Concreteness スコアは 1 から 5 までの連続値で，約 4000 単語に対してこのスコアが与えられている．このスコアを用いて，VSNLI の dev と test の前提文において 10 回以上出現する単語のうち[2]，concreteness スコアが付与されている単語を選んで，ダミーの特徴量への注意の重みと concreteness スコアとの相関をとった．ダミーの特徴量は単語と画像の非対応関係を捉え，画像中に対応する部分のある具体的な概念を表す単語からはダミーへの注意の重みが小さく，抽象語や機能語からはダミーへの注意の重みが大きくなるという想定であったので，これが正しく実現されていれば，ダミーの特徴量へ注意の重みと concreteness スコアとの間には負の相関が見られることが期待される．

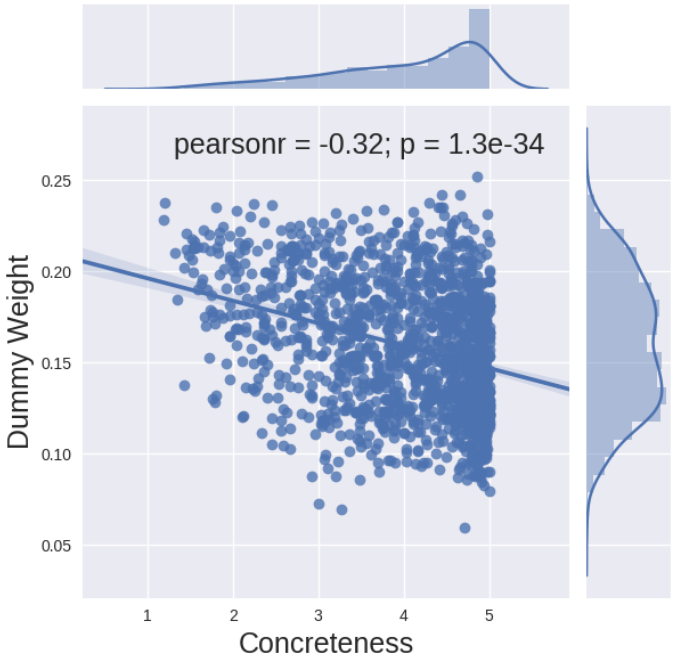
図 2 にこの結果を示す．図 2 上が元の画像を使った場合の相関，図 2 下は入れ替えた画像を使った場合の相関である．図 2 から，元の画像を使った場合にはダミーの特徴量への注意の重みと concreteness スコアには-0.32 の負の相関があり，仮説のとおりダミーの特徴量はクエリの単語と画像との非対応関係を捉えていると考えられる．入れ替えた画像を用いた場合はこれが-0.26 になり，画像入れ替え前よりやや無相関に寄っている．このことから，元画像を使った場合は抽象語や機能語など concreteness スコアの低い単語がダミーへの注意の重みを大きくしているが，画像入れ替えによって単語が画像に対応する部分をもたなくなると，concreteness スコアの高い具体的な概念を表す単語が画像中での対応を失ってダミーへの注意の重みを大きくしてしまっていることが示唆される．

具体的にこれを確かめるため，表 7 において，画像の入れ替えによって解答を誤るようになる設問について，ダミーへの注意の重みと concreteness スコアを示した．この表から，前提文中の “man”，“road” といった元々画像に対応のあった具体的な概念を表す単語が，画像の入れ替えによって画像中に対応をなくし，相対的にダミーへの注意の重みを大きくしていることが見てとれる．図 2 では仮説

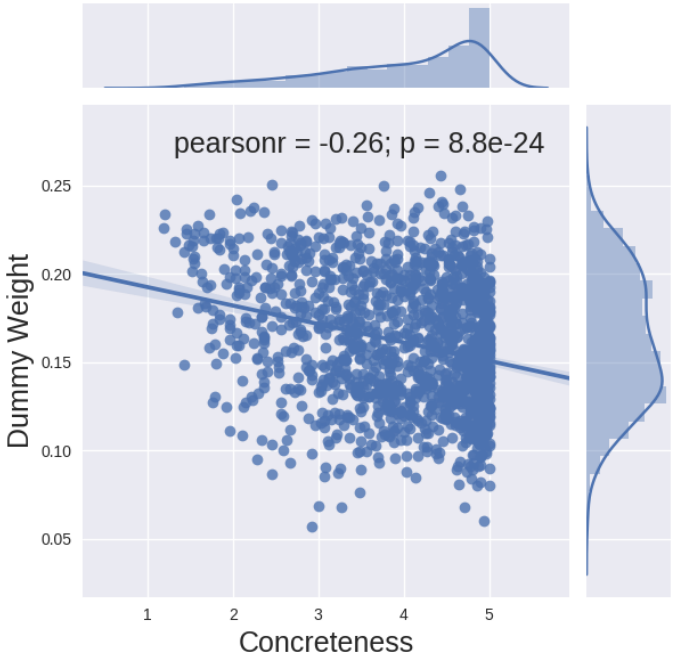
[2]仮説文中には具体的な概念を表す単語であっても元々与えられた画像と対応しない単語が含まれるため，ここでは前提文のみを用いた．

文中の単語については扱っていないが，表 7 では解答の誤りの原因の分析のために仮説文でのダミーへの注意の重みも示してある．仮説文中の単語では “buying”, “lottery”, “ticket” が具体的な概念を表す単語であるが，“man”, “road” とは逆の理由で，画像の入れ替えによる判断の誤りが見られる．これらは元の画像では明らかに対応する部分をもたなかったため，具体的な概念を表す単語であるのにダミーへの注意の重みが大きかったが，画像を入れ替えたことでこの明らかな非対応の関係をなくし，画像入れ替え後逆にダミーへの注意の重みを小さくしてしまっている．この設問には 6.1 節で見た Entity と Verb の知識が必要とされており，具体的には，仮説文中で示されている具体的な物 (lottery，ticket) と具体的な行為 (buying) が，前提文で表現されている内容と異なるということがわからなければならない．これらの単語が画像の入れ替えによって上記のような誤った注意の重みを与えてしまっているため，画像の入れ替えによって解答を誤るようになったと考えられる．参考のため，図 3 にこの設問で扱った画像と注意の重みの可視化の結果を示す．

以上から，ダミーの特徴量への注意の重みはクエリの単語と画像との (非) 対応の度合いを表していると考えられる．前節までの考察から提案手法での精度の向上はこの (非) 対応の度合いによる単語特徴量の重み付けのみによるので，提案手法では，画像特徴量は具体的な物や行為を表す単語との対応の度合いとして，限定的な活用をされていると考えられる．



(a) Original image

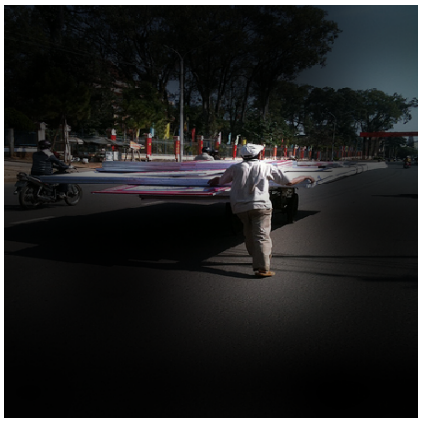


(b) Foil image

図 2: ダミーへの注意の重みと concreteness スコアの相関

	Query	Original dummy value	Foil dummy value	Concreteness
Hypothesis	a	0.23	0.23	1.46
	man	0.19	**0.21**	**4.79**
	is	0.22	0.23	1.59
	buying	**0.20**	0.19	**(buy 3.35)**
	a	0.23	0.23	1.46
	lottery	**0.14**	0.11	**4.19**
	ticket	**0.15**	0.14	**4.7**
	.	0.23	0.24	-
Premise	a	0.23	0.23	
	man	0.19	**0.21**	**4.79**
	,	0.25	0.25	-
	wearing	0.22	0.22	3.59
	a	0.23	0.23	1.46
	cap	0.14	0.13	4.59
	,	0.25	0.25	-
	is	0.22	0.23	1.59
	pushing	0.15	0.16	3.73
	a	0.23	0.23	1.46
	cart	0.12	0.10	4.89
	,	0.25	0.25	-
	on	0.21	0.22	3.25
	which	0.22	0.21	1.54
	large	0.15	0.12	3.37
	display	0.15	0.15	3.86
	boards	0.15	0.15	(board 4.57)
	are	0.21	0.22	1.96
	kept	0.14	0.11	2.79
	,	0.25	0.25	-
	on	0.21	0.22	3.25
	a	0.23	0.23	1.46
	road	0.13	**0.14**	**4.75**
	.	0.23	0.24	-

表 7: 画像の入れ替えによって分類を誤る例における，ダミーへの注意の重みと concreteness スコア．Query は注意機構でのクエリの単語，Original dummy value は元々与えられた画像でのダミーへの注意の重み，Foil dummy value は入れ替えられた画像でのダミーへの注意の重みである．



(a) Original: man

(b) Foil: man

(c) Original: lottery

(d) Foil: lottery

(e) Original: buying

(f) Foil: buying

図 3: 注意の重みの具体例．画像中で明るい部分ほど，注意の重みが大きことを示す．Original は元々与えられていた画像，Foil は入れ替えられた画像．Original/Foil に続く単語は注意機構におけるクエリである．

7. おわりに

7.1 結論

本研究では，注意機構によって画像中から単語に対応した部分を選択することで，より詳細な言語表現と画像の対応付けと，その分析を行うことを目的とした．注意機構にダミー特徴量付きの注意機構を用いることで単語から画像への対応/非対応関係の両面を考慮した画像特徴量選択を行い，画像中に対応する部分をもたない単語に対してはその非対応の度合いに応じて単語特徴量を補完する単語-画像特徴量ゲートを提案した．実験の結果，提案手法での画像情報の導入によって含意関係認識の精度が向上することが確認されたが，分析では，画像特徴量は具体的な物や行為を表す単語との対応の度合いとして限定的にしか用いられていないことがわかった．

7.2 今後の課題

本研究では，画像特徴量がネットワーク中で限定的にしか用いられないという課題が見つかった．この原因のひとつは，注意機構による画像特徴量選択を含意関係認識のタスクと同時に学習していることにある可能性がある．InferSent のベースラインが高い精度を示していることからわかるように，含意関係認識は単語特徴量だけである程度解けることから，このタスクでは，画像特徴量について十分に学習を行わずに単語特徴量だけを使ってしまう局所解に陥りやすいことが考えられる．このことへの対策として，注意機構を他のタスクで事前学習してから含意関係認識に転移学習することを考えている．これを今後の課題としたい．

謝辞

研究会や勉強会で様々に指導，助言してくださった，松本裕治教授，新保仁准教授，進藤裕之助教に感謝致します．副査の中村哲教授には，ゼミナールの中間発表で有益なコメントをいただきました．秘書の北川祐子さんには，研究室での生活面で大変お世話になりました．研究室では優秀な同期，先輩，後輩に恵まれました．修士の 2 年間は外部の研究所の方々にもメンターとなっていただき，助言をいただきました．NTT コミュニケーション科学基礎研究所の平尾努さん，オムロンサイニックエックス株式会社の牛久祥孝さん，橋本敦史さんに感謝申し上げます．最後に，これまで支えてくれた家族に感謝を伝えたいと思います．

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[3] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 2014.

[4] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *ACL*, 2018.

[7] Dan Han, Pascual Martínez-Gómez, and Koji Mineshima. Visual denotations for recognizing textual entailment. In *EMNLP*, 2017.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[9] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. In *NAACL-HLT*, 2018.

[10] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *ICLR*, 2017.

[11] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Bilinear attention networks. In *NIPS*, 2018.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[13] Jamie Kiros, William Chan, and Geoffrey Hinton. Illustrative language understanding: Large-scale visual grounding with image search. In *ACL*, 2018.

[14] Alice Lai and Julia Hockenmaier. Illinois-lh: A denotational and distributional approach to semantics. In *SemEval*, 2014.

[15] Alice Lai and Julia Hockenmaier. Learning to predict denotational probabilities for modeling entailment. In *EACL*, 2017.

[16] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, 2014.

[17] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014.

[18] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*, 2018.

[19] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[20] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 2017.

[21] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *ICLR*, 2018.

[22] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[26] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

[27] Hoa Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. In *COLING*, 2018.

[28] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. 2017.

[29] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *ICLR*, 2016.

[30] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[31] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014.

発表文献リスト

Ukyo Honda, Tsutomu Hirao, and Masaaki Nagata. Pruning basic elements for better automatic evaluation of summaries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 661–666. Association for Computational Linguistics, 2018.

本多右京，平尾努，永田昌明．冗長な Ngram/Basic Elements を除いた自動要約評価指標．言語処理学会第 24 回年次大会発表論文集，pp.17-20, 2018.