テキストを通して世界を見る: 機械読解における常識的推論のための画像説明文の評価

Diana Galvan-Sosa[1] 西田 京介 [2] 松田 耕史 [3,1] 鈴木 潤 [1,3] 乾 健太郎 [1,3]
[1] 東北大学 [2]NTT メディアインテリジェンス研究所 [3] 理化学研究所
{dianags, matsuda, jun.suzuki, inui}@ecei.tohoku.ac.jp
kyosuke.nishida.rx@hco.ntt.co.jp

1 はじめに

質問応答 [14] や機械読解 [7] などの自然言語理解タスクにおける最近の進歩は、自然言語処理 (NLP) コミュニティから注目を集めている。大量のテキストコーパスでモデルを訓練することで，質問応答や機械読解をはじめとした NLP タスクを解くために必要な，さまざまな知識をモデルに導入可能である．しかし，これはどんな種類の知識にも当てはまるだろうか？我々は，大量のテキストコーパスのみからは獲得が難しいタイプの知識も存在すると考える．たとえば，**常識的知識**は，現実世界のさまざまな相互作用に基づいて時間をかけて発展していくという特殊性を持っている．そのため，常識的な知識は，テキストの形で明示的に書かれることはまれである．近年では，ConceptNet [9] や Event2mind [15] や Atomic [17] など、テキストコーパスだけを用いるのではなく，クラウドソーシングされた注釈を集めることでこの問題を軽減する試みがいくつかある．しかし，これらのアプローチにおいては，獲得される知識表現の型が（主語・述語・目的語）からなるトリプルに限定されており，トリプルの述語の語彙も閉じているため，自然言語での記述に比べて知識の表現力という面で劣る．さらに，アノテーターに問い合わせる知識の骨格となる部分の多くを既存のテキストに依拠しているため，現実世界では真実であるが，テキスト中にはあまり出現しない知識の多くを見過ごしている．

本研究においては，理解とは、Schank や Abelson が述べたように，人々が**見たり**聞いたりしたことと、すでに経験したことを一致させるプロセスであるという前提に立つ [18]. この前提にの上に立つと，視覚入力に対してなされたアノテーション（例えば，画像の説明文など）には，書き言葉に比べて常識的な知識が多く含まれると予想される．

我々の最終的な目標は，コンピュータビジョンに関する研究分野で構築されてきた画像説明文に関するデータセット（または，キャプション生成等のモデルからあらたに生成された説明文）から獲得された知識を用いて，与えられた状況に関する常識的な質問に答えることができるシステムを構築することである．そのためにまず，常識的な質問からなる質問応答タスクにおいて，画像説明文から獲得された知識の評価を行うことから始める．本研究の貢献は以下の通りである：

- 画像説明文の最も大きく高密度なデータセットである Visual Genome [6] に含まれる常識的な知識を抽出し，機械読解タスクを通してその評価を行うフレームワークを提案する．
- 上述のフレームワークによって検索された画像説明文を用いることで，事前訓練された BERT [2] がどのように fine-tuning されるのかを調査する．

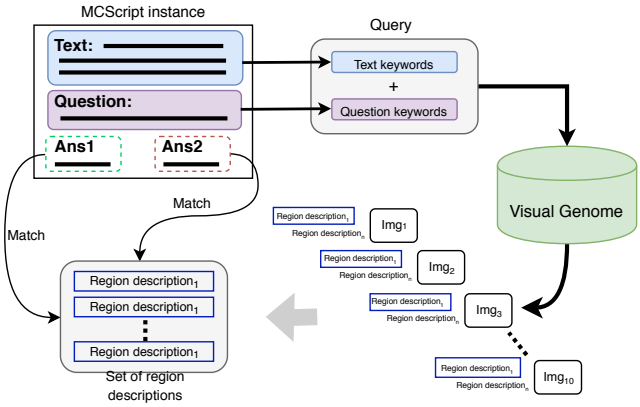


図1: 提案するフレームワーク．MCScript 機械読解タスクにおける質問と文脈から生成したクエリで Visual Genome を検索し，検索された画像領域に対する説明文と解答候補とのマッチングを行う．

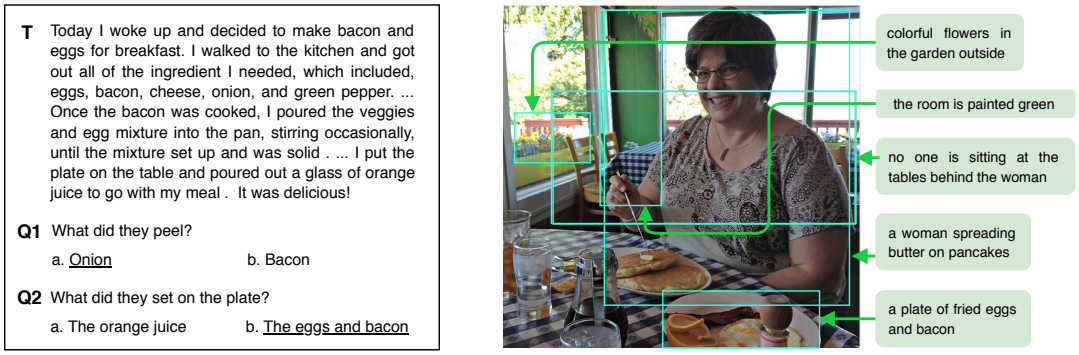
2 関連研究

知識獲得. 画像説明文データセットが，常識知識抽出にどのように役立つかを調べた研究はいくつかある．例えば，Yatskar ら [20] および Mukuze ら [10] は、MS-COCO データセット [8] から 16K の常識関係および 2K の動詞/場所の対 (e.g., *holds*(dining-table, cutlery), *eat*/*restaurant*) を導出した．しかし，彼らは物理的な関係からなる知識にのみ焦点を合わせている．

もう一つの最近のトレンドとして，言語モデルに常識的な知識を問い合わせる研究が存在する． BERT のような堅牢な言語モデルは、常識的知識を事実的知識と同程度のレベルで検索可能な強力な性能を示しているが [13]、実際にはその知識が明示的にテキストコーパスに存在する場合にのみ問い合わせ可能であることが指摘されている [3].

機械読解. 機械読解タスクは，与えられたテキストに対する質問を通して，計算機の言語理解度を評価するタスクである． Visual Question Answering [4], MCScript [11, 12], Visual Commonsense Reasoning [21] および CosmosQA [5] のような現在の最も挑戦的なデータセットは、文脈 (テキストや画像の形で与えられる) と背景知識の両方を使用することによってのみ解けるように設計されている．これらのデータセットにおいては，すでに提案されているいかなるシステムも，人間の正答率に及んでいないのが現実である．このことは，システムが利用可能な知識源を新たに見つける必要があることを示唆する．

我々の研究は，この二つの方向性が交差するところに位置づけられる．我々は，広範な常識知識獲得のために，画像に（手



(a) MCScript instance

(b) Visual Genome instance

図2: (a) MCScript の例: 日常的な活動に関する物語テキストと，常識的知識を要する (Q1) またはテキストからの推論で解答可能な (Q2) いくつかの読解問題から構成される．(b) Visual Genome データセットの例: 多数の領域アノテーションと，領域に付与された説明文から構成される．

動で・あるいは将来的には自動的に）付与されたテキストを活用するという方向性について論じる．また，Visual Question Answering や Visual Commonsense Reasoning タスクにおけるマルチモーダル情報の活用に関する取り組みの成功を，機械読解タスクにおいても達成するための試みという位置づけも可能である．

3 データセット

本研究で取り扱う 2 つのデータセットの例を図 2 に示す．本節では，それらの概要について紹介する．

MCScript & MCScript 2.0 [11, 12] は，日常的な活動（たとえば，TAKING A SHOWER や MAKING BREAKFAST など）に関する短い物語を題材とした読解データセットである．読解対象となる各テキストには質問がいくつか付与されており[*1]，それぞれの質問には，二つの解答候補が付与されている．一方は正しく，他方は間違った解答候補である．データセット中のそれぞれの質問は次のラベルのいずれかが付与されている[*2]．

- **Text:** テキスト中に記述されている知識のみから解答が可能である質問
- **Commonsense:** テキスト中に明に記述されない常識的な知識を用いることで初めて回答可能となる質問

MCScript と MCScript 2.0 にはそれぞれ 13,939 と 19,821 の質問が含まれている．実験を行うにあたり，あらかじめ，Osterman ら [11, 12] と同様の手順にしたがって，データを訓練・開発・評価セットに分割した．

Visual Genome [6] は，画像内の複数のオブジェクトおよび領域に対して高密度な説明文を持つ大規模なデータセットである．画像に対して単一の説明文が付与されている他のデータセットとは対象的に，Visual Genome においては画像中に出現するいくつもの主要なオブジェクトやシーンに対して，領域情報とその領域に関する説明文が付与されている．データセット内の約 108,000 の画像それぞれには平均 50 の領域情報が付与され，それぞれの領域に最大 16 ワードの説明文が付与されている[*3]．本研究では，Visual Genome に付与された説明文アノテーションを知識源として読解タスクを解くことを通して，一般的なテキストコーパスから得られる情報に加えて視覚に依拠した情報を自然言語処理モデルに組み込むことについて議論する．

4 提案手法

我々の仮説は，大規模画像データセットに付与された注釈（領域に付与された説明文）が，常識的知識を明示的に例示している，ということである．この仮説に立つならば，それらの明示的に表現された知識を用いて，言語処理システムの性能を改善できるはずである．この仮説について検証するために，日常的なシナリオに関する読解データセットである MCScript の質問に答えるために，Visual Genome における画像の高密度な説明文がどの程度十分であるかを調査した．

我々の提案するフレームワークを図 1に図示する．MCScript にはテキスト・質問・二つの解答候補という要素が含まれる．まず，テキストと質問からキーワードを抽出し，クエリを生成する．次に，キーワードを連結した上で Visual Genome に対してクエリを実行し，上位 10 件の画像検索結果と，それらに付随する領域説明文を得る．また，それぞれの領域説明文と解答候補の類似性スコアを計算した．最終スコアは解答候補とすべての領域説明文の類似度の平均である．このスコアを各回答候補に対して計算し，よりスコアが高い解答候補を出力する．

4.1 キーワード抽出と知識検索

開発セット内の各テキスト – 質問対からキーワードを抽出するために，TF-IDF を使用して，テキストと質問のそれぞれからスコア上位の最大 10 個のキーワードを抽出した[*4]．抽出されたキーワードを連結してクエリを生成して Visual Genome に問い合わせ，上位 10 件の画像検索結果を得る．画像検索結果から，クエリと領域説明文の一致度に基づいて，最大 3 件の領域を抽出した．検索された領域説明文の集合と各回答候補の類似度尺度にはコサイン類似度を用いた．

[*1] 平均すると、MCScript においてはテキストあたり 6.7 問、MCScript2.0 においてはテキストあたり 5.7 問の質問が付与されている．

[*2] MCScript 2.0 には，それらに加えて”positive-merged” というラベルが付与された質問も存在する．これは，解答のためにテキスト中に明に記述されている知識または常識的な知識のどちらかが必要な問題である．本論文の後半部では，”Text/Commonsense” というラベルで参照する．

[*3] 実験においては，Elasticsearch https://www.elastic.co/を用いて各画像の領域に対する説明文を索引付けした．

[*4] IDF の計算には両バージョンの MCScript のトレーニングデータのテキストを結合したものを用いた．また，キーワード抽出前に stop words を除去している．

Data	MCScript			MCScript2.0			
Model	Text	Commonsense	Total	Text	Commonsense	Text/Commonsense	Total
最新のモデル							
TriAN (Wang et al., 2018) [19]	-	-	85.27	-	-	-	-
HFL-RC (Chen et al., 2018) [1]	-	-	86.46	-	-	-	-
類似度に基づくベースライン							
TF-IDF vectors	52.4	50.4	51.8	53.8	54.2	55.2	54.2
SBERT embeddings	55.6	**56.0**	**55.7**	**58.8**	**60.5**	**59.5**	**59.7**
SBERT embeddings (all regions)	**55.8**	54.1	55.3	50.4	55.0	53.8	52.9
Fine-tune された BERT							
vanilla-BERT	88.27	82.72	86.68	**86.02**	75.16	75.71	79.75
Visually Enhanced BERT	**88.37**	**83.70**	**87.03**	85.31	**77.74**	**76.77**	**80.79**

表1: 両バージョンの MCScript 開発セットにおける，提案手法の正解率を示す．最上段は最近提案されたモデルの正解率である．

5 実験

5.1 類似度に基づくベースライン

まず，解答候補と画像領域説明文の意味的類似性に関するテストを行った．領域説明文が常識的な情報を持っている場合，コサイン類似度は正しい回答候補に対してより大きく傾向があるはずである．このことを確かめるために，TF-IDF ベクトルと BERT 文埋め込み (SBERT) [16] を用いて，文と解答候補のベクトル表現を得た．各回答候補の類似性スコアを得た後、より高いスコアの候補を選択した．同点の場合は、ランダムに 1 つの回答を選択した．

5.2 BERT の Fine-tune

vanilla-BERT: 事前訓練された BERT を次の入力構成で Fine-tune した．質問とその回答候補の一つを連結したトークン列をセグメント 1 に追加し，テキストのトークン列をセグメント 2 に追加する．もう一つの回答候補についても同様に系列を作成する．結果として，インスタンスごとに 2 つの系列が BERT に入力され，それぞれの分類用トークン [CLS] に対応する状態ベクトルを 1 次元に線形変換して，softmax により回答を決定する．セグメント 1 と 2 を分けるセパレーター・トークン [SEP] と別に，BERT がセグメント 1 内の質問と回答候補トークンを区別できるように，特別なセパレーター・トークンを使用した．学習においてはクロスエントロピー最小化を行う．

Visually Enhanced BERT: 4.1の検索結果を使用して，Visual Genome から一連の領域説明文を取得する．vanilla-BERT からの変更は，セグメント 2 においてテキストのトークン列と検索された領域説明文のトークン列を特別なセパレータ・トークンで連結した点である[*5]．我々の目標は，領域説明文を用いて，特定のシーンやシナリオに関する背景知識を含んだ機械読解モデルを構築することである．このため，入力に対する領域説明文の連結は訓練時にのみ行った．評価時には，インスタンスを vanilla-BERT と同様に扱うことによって，学習時に Visually Enhanced BERT が獲得した背景知識についての検証を行う．

6 結果

6.1 類似度に基づくベースライン

表 1に，類似度に基づくベースラインの結果を示す．TF-IDF に基づくベクトル表現は MCScript に対しては，特に Commonsense 問題において，ランダムに選択した場合と殆ど変わらない正解率であった．MCScript 2.0 に対する正解率に着目すると，Text 質問より Commonsense 質問に対する正解率が上回っていた．この低い性能の原因は文の意味表現が正当に計算できていないことにに起因する可能性について考慮し，我々は BERT に基づく文表現計算手法の一つである SBERT を用いて文の埋め込みを計算した．SBERT は両方のデータセットに対して高い性能を発揮した．特に Commonsense 問題，Text/Commonsense 問題において顕著に性能が向上した．

最後の実験として，SBERT 埋め込みにおいて，4.1節で紹介した検索の方法を修正し，検索結果上位 10 件の画像のすべての領域説明文を類似度計算に用いることを試みた．この設定においては，より多くの視覚的な情報を用いることができるにもかかわらず，両方のデータセットにおいて性能の向上がみられなかった．このことは，画像全体が質問に関係していない場合，説明文が逆にノイズになる可能性を示唆している．たとえば，*What was packed to take pictures?* という質問に対して，*silver camera to take pictures* という領域説明文がマッチした．この画像には，他にも *A logo is on the red panel of an umbrella*, *Man wearing floral jacket* のような領域説明文が付与されており，これらは質問に対する解答を推論する上でノイズになりえる．

6.2 Fine-tune された BERT

vanilla-BERT ベースラインは，類似度に基づくベースラインよりも良好に機能した．しかし，Text 問題に対する際立った性能と比較すると，Commonsense 問題，および Text/Commonsense 問題における性能においては一歩譲る．そこで我々は，本稿で提案したフレームワークにより検索された領域説明文を用いて Commonsense 問題に対する vanilla-BERT の性能を更に改善することを目指した（Visually Enhanced BERT）．表の最下部に見られるように，領域説明文を用いて Fine-tune されたモデルは MCScript における正解率，および MCScript 2.0 における Commonsense 問題と，Text/Commonsense 問題の正解率を

[*5] セグメント 1 とセグメント 2 のための特別なセパレーター・トークンとして，BERT の語彙に含まれる [unused00] と [unused01] を使用した．

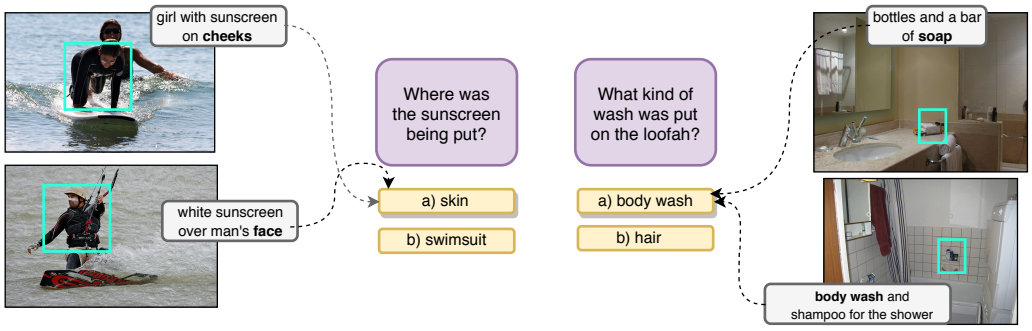


図3: MCScript 2.0 より，二つのサンプル質問を示す．これらは BERT では正しい答えを選択することができなかったが，提案フレームワークによって検索された画像領域説明文を使用して、Visually Enhanced BERT は正しい答えを選択することが可能になった．

向上させた．このことは，Visual Genome の領域説明文が常識的な知識の源として利用可能であることを示唆している．

我々の仮定の妥当性をさらに検証するために，vanilla-BERT によって正しく答えられなかった Commonsense 関連問題のいくつが，類似度に基づく最も強いベースライン（SBERT）によって正しく解答できたかを，MCScript 2.0 に対して確認した． その結果，vanilla-BERT が誤答した 241 の質問のうち，SBERT は 111 の質問に対して正しく解答できていたことを確認した．このことは、質問のタイプを分類することが可能になれば，Commonsense 質問に対して我々のアプローチを使用することが有効であることを示唆する．

6.3 ケーススタディ

図 3は、VISITING THE BEACH および TAKING A SHOWER シナリオからそれぞれ、質問 *Where was the sunscreen being put?* および質問 *What kind of wash was put on the loofah?* に対して検索された画像の領域とその説明文を示している．大規模なコーパスで事前に訓練されていたにもかかわらず, BERT は，*sunscreen* と *skin* との間の関係を推論することができず，*swimsuit* を正しい答えと判断した．同様に，*loofah* と *body wash* の関連付けにも失敗している．対して Visually Enhanced BERT においては，*white strip of sunscreen on man's face* という説明が付与された領域が存在したため，正解することができた．

第 2 の質問では、領域説明文に *loofah* の明示的な記述はなかったが、いずれも *body wash* との関連性が高く，検索した領域説明文で vanilla-BERT を補完したところ、正しい回答が得られた．これまでの調査では、*what* と *where* 質問に対して画像から知識を獲得することは，それが画像とその説明文に記述されることが多いため，合理的に見える．しかし，その他のタイプの質問が画像や動画などの視覚的な情報を用いることで改善されるのかについては，今後の調査が必要である．

7 おわりに

本稿では，密な画像説明文から知識を得ることの妥当性を検討した．常識的知識が必要な機械読解タスクにおける予備的結果は，画像領域の説明文が，通常のテキストには見られない，単純だが価値ある情報を含むことを示唆した．今後は、画像とその説明を含む他の大規模データセットに手法の適用範囲を拡張し，CosmosQA のようなより挑戦的な機械読解データセットで評価することを検討している．

謝辞: 本研究の一部は，JST CREST(課題番号: JPMJCR1513)，および，JSPS 科研費 JP19H04162 の支援を受けたものである．

参考文献

[1] Zhipeng Chen et al. “HFL-RC system at SemEval-2018 task 11: hybrid multi-aspects model for commonsense reading comprehension”. In: *arXiv preprint arXiv:1803.05655* (2018).
[2] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
[3] Maxwell Forbes, Ari Holtzman, and Yejin Choi. “Do Neural Language Representations Learn Physical Commonsense?” In: *arXiv preprint arXiv:1908.02899* (2019).
[4] Yash Goyal et al. “Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering”. In: *CVPR*. 2017.
[5] Lifu Huang et al. “Cosmos QA: Machine reading comprehension with contextual commonsense reasoning”. In: *arXiv preprint arXiv:1909.00277* (2019).
[6] Ranjay Krishna et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *IJCV* (2017).
[7] Guokun Lai et al. “Race: Large-scale reading comprehension dataset from examinations”. In: *arXiv preprint arXiv:1704.04683* (2017).
[8] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *ECCV*. Springer. 2014.
[9] Hugo Liu and Push Singh. “ConceptNet: a practical commonsense reasoning tool-kit”. In: *BT technology journal* 22.4 (2004), pp. 211–226.
[10] Nelson Mukuze et al. “A vision-grounded dataset for predicting typical locations for verbs”. In: *LREC*. 2018.
[11] Simon Ostermann, Michael Roth, and Manfred Pinkal. “MCScript2. 0: A Machine Comprehension Corpus Focused on Script Events and Participants”. In: *arXiv preprint arXiv:1905.09531* (2019).
[12] Simon Ostermann et al. “Mcscript: A novel dataset for assessing machine comprehension using script knowledge”. In: *arXiv preprint arXiv:1803.05223* (2018).
[13] Fabio Petroni et al. “Language Models as Knowledge Bases?” In: *arXiv preprint arXiv:1909.01066* (2019).
[14] Pranav Rajpurkar et al. “Squad: 100,000+ questions for machine comprehension of text”. In: *arXiv preprint arXiv:1606.05250* (2016).
[15] Hannah Rashkin et al. “Event2mind: Commonsense inference on events, intents, and reactions”. In: *arXiv preprint arXiv:1805.06939* (2018).
[16] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *EMNLP*. 2019.
[17] Maarten Sap et al. “Atomic: An atlas of machine commonsense for if-then reasoning”. In: *AAAI*. 2019.
[18] Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013.
[19] Liang Wang et al. “Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension”. In: *arXiv preprint arXiv:1803.00191* (2018).
[20] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. “Stating the obvious: Extracting visual common sense knowledge”. In: *NAACL*. 2016.
[21] Rowan Zellers et al. “From recognition to cognition: Visual commonsense reasoning”. In: *CVPR*. 2019.