言語と画像の統合理解のための文書への画像挿入タスク

石井悦子 †1*　　小比田涼介 ‡2　　村岡雅康 ‡3

† 東京大学 工学部計数工学科
‡ 日本アイ・ビー・エム株式会社 東京基礎研究所
[1] ishii@sat.t.u-tokyo.ac.jp [2] kohi@ibm.com [3] mmuraoka@jp.ibm.com

1 序論

近年，自然言語処理と画像処理を融合したマルチモーダル領域の研究が盛んに行われており，深層学習の発展により，画像のキャプション生成 [12] やクロスメディア検索 [3, 6] といった技術の進歩が目覚ましい．これらの技術に共通しているのは，テキストと画像のような異なるデータ形式を比較可能な表現にエンコードし，内容の意味的な類似度を測る技術が重要である．しかし既存手法では，基本的には 1-2 文と画像のペアで学習が行われるものが多く，大きな文脈を考慮することができないと考えられる．そこで本稿では，文章全体を考慮し文章中の適切な位置に適切な画像を挿入する「画像挿入タスク」を新たに提案する．

SNS や動画像共有サービスの普及により，マルチモーダルデータが爆発的に増加する中，画像挿入タスクは次のような応用が考えられる．SNS の一つである Facebook[1]や，全国の飲食店の情報が掲載されているグルメサイト「食べログ」[2]では，ユーザが書いた文章に画像を添付して投稿できる．しかし文章と画像は切り離されて表示される場合もあり，閲覧ユーザは文章と画像を個別に対応付けながら文章を読まなければならない．これを自動で対応付けることができれば可読性が大きく向上するだろう．また，作文された文章に適切な画像を自動で挿入することができれば，SNS の投稿やフォトアルバム作成の支援が期待できる．

これを実現するためには，文と画像の類似度学習だけでなく，文章構造や段落間・画像間の関連性も考慮する必要がある．マルチモーダルデータを用いてこのような高度な意味内容理解を問う手法・課題は，我々の知る限り存在しない．本稿では，文書への画像挿入を行う課題及びデータセットを提案し，全体として最適な位置に挿入を行う手法を考案，その性能評価を行う．

*この研究は日本アイ・ビー・エム株式会社 東京基礎研究所のインターンシップとして行われたものである．

[1] https://www.facebook.com/

[2] https://tabelog.com

2 関連研究

ここでは，本稿に関連のあるマルチモーダル領域の研究について述べる．

自然言語文をクエリとして，その内容を表す画像を抽出したり，画像をクエリとして，適切なキャプションを獲得するタスクをクロスメディア検索 [11] と呼ぶ．代表的な手法として，CCA(Canonical Correlation Analysis) や DNN(Deep Neural Network) を用いて，テキストと画像のそれぞのれ特徴量空間から共通空間へのエンコーダを学習するものがある [3, 5, 8]．共通空間へのエンコーダは，意味的内容が同じテキストと画像は共通空間において互いに近くなるように学習される．

一方で，MS COCO[9] や Flickr 30k[15]，VQA[4] といったデータセットが開発されたことで，入力画像を説明するキャプション生成タスクや，入力画像に関する質問に答える Visual Question Answering 技術が飛躍的に進歩した．このようなタスクでは，入力となる画像や質問文をエンコードするエンコーダと，エンコードされたエンベディングを受け取り，テキストを生成するデコーダが学習される [10, 12]．しかしながら，これらはいずれも 1-2 文程度のテキストと画像を用いて学習するため，文章構造のようなより高度な意味内容を理解する学習は行われない．

こういった背景から，文脈を考慮しつつ文章-画像を結びつけるユニークなタスクが近年いくつか考案されている．とりわけ目を引くのが，時系列を “文脈” として取り入れたタスクである．例えば，4 コマ漫画の起承転結のように，キャプション付き画像を時系列順に並べかえるもの [1] や，漫画のコマとコマの間の飛躍の理解度を試すもの [7]，フォトアルバムを用いてあるシナリオにおいて典型的なイベントの流れ (例えば，結婚式は，新郎新婦の入場→指輪交換→ケーキカット等の流れがある) を学習するもの [2] があげられる．

3　提案タスク

この章では提案タスクの問題設定を定義し，既存タスクと比較して特徴的な側面を挙げる．その上で，具体的なデータの構築方法について触れる．

3.1　問題設定

提案タスクの問題設定は以下のように定式化される．入力文書 D は，N 個のセクション $s_{1:N} = s_1, \cdots, s_N$ からなるセクション集合 S と M 個の画像 $p_{1:M} = p_1, \cdots, p_M$ からなる画像集合 P によって構成され，画像 p はいずれか一つのセクション s に帰属する．セクション s は一つ以上の文からなり、多くの場合三文以上である．提案タスクのゴールは、文書 D の全ての画像 p_k を正しいセクション $\hat{s}_k$ に割り振ることである．

3.2　タスクの特徴

既存タスクと比較して、本タスクでの特徴的な点として以下が挙げられる．

(a) 画像との関係性を捉える対象（＝セクション）が複数文からなる　あるセクションに対する画像の挿入の有無は、当該セクション全体の意味に依存する．多くの文章-画像タスクでは 1 画像に対して 1·2 文程度を扱うが、本タスクでは複数文、場合によっては数十文が対象となるため，複数文からの内容理解や文書要約といった能力が要求される．

(b) セクションが互いに関連性を持つ　セクションは互いに構造的（e.g. 順序，階層），及び，意味的（e.g. 起承転結，因果関係）に関連性を持ち，その関連性が画像の挿入位置に影響を与える．文書中のセクション同士の関係性は抽象的で，十分な読解能力が要求される．

(c) 画像が互いに関連性を持つ　他の画像の挿入位置に依存して自身の位置が決まるものや，複数の画像によって一つの意味を成すものなど，画像も互いに関連を持ちうる．複数画像からの意味抽出には，その背景となる知識や前提の理解が重要であり，文書と照らし合わせて推論する能力が要求される．

このように，本タスクは文書側も画像側も高度な理解と推論が要求されるものである．これは，クロスメディア研究におけるタスクに留まらず，機械による意味理解という課題に対するベンチマークとしても有用であることを示唆している．

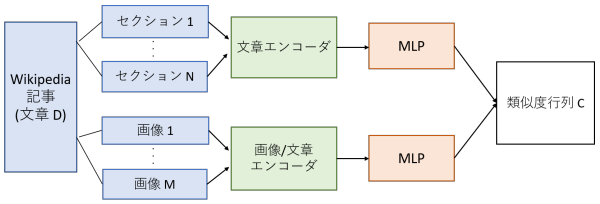

図 1: セクション-画像間の類似度行列 C の算出モデル．エンコーダは学習済みのものを使用し，MLP 部分のみ学習する．

3.3　データ構築

提案タスクのデータは Wikipedia 記事を対象に，HTML タグの h タグで囲まれている部分をセクションとみなすことで，自動構築することができる．文書 D が記事全体，セクション s が h タグで囲まれている文，画像 p が記事内の各画像にそれぞれ対応し，画像 p_k が挿入されていたセクションが $\hat{s}_k$ となる．

Wikipedia から自動構築できる点も本タスクの利点であり，大規模データ，一般・ドメイン知識，多言語モデル，最新知識の拡充といった重要事項を満たすデータセットを，コストをかけずに構築できる．

4　提案手法

本稿では，画像の挿入箇所として相応しいセクションは意味的に近くあるべきと考え．画像と挿入セクションの類似度をもとに挿入箇所を定める手法を提案する．まず，セクションと画像をそれぞれベクトル表現に変換後（式 1, 2），ペアワイズで類似度を計算し，類似度行列 $C = (C_{i,k}) \in \mathbb{R}^{N \times M}$ を獲得する（式 3）．

$$v_{s_i} = E_s(s_i),\ v_{p_k} = E_p(p_k) \tag{1}$$

$$h_{s_i} = MLP_s(v_{s_i}),\ h_{p_k} = MLP_p(v_{p_k}) \tag{2}$$

$$C_{i,k} = \text{Similarity}(h_{s_i}, h_{p_k}) = \cos(h_{s_i}, h_{p_k}) \tag{3}$$

E_s, E_p は文章エンコーダと画像エンコーダで，それぞれセクション s_i を l 次元ベクトル v_{s_i} に，画像 p_i を r 次元ベクトル v_{p_k} に変換する．MLP_s, MLP_p は多層パーセプトロンであり，文章ベクトル v_{s_i} と画像ベクトル v_{p_k} を q 次元の h_{s_i}, h_{p_k} に写像する．画像 p_k のキャプションを利用する場合，E_s で変換した $v^{\text{cap}}_{p_k}$ と v_{p_k} を結合したベクトルを MLP_p の入力とする．

$C_{i,k}$ は画像 p_k のセクション s_i に対するスコアであるが，記事全体を考慮できていない．そのため $C_{i,k}$ ではなく，式 4 で定義される線形計画問題の解：確率行

列 $X = (X_{i,k}) \in \mathbb{R}^{N \times M}$ を最終スコアとして用いる．$X_{i,k}$ はセクション i に画像 k が挿入される確率を表しており，最も確率の高いセクションが p_k の予測挿入箇所 $s_k^{'}$ となる．

$$\max \sum_{i,k} C_{i,k} X_{i,k} \quad \text{s.t.} \quad \sum_{i} X_{i,k} = 1. \tag{4}$$

学習時は，式 4 から得た解 X と真の解 $\bar{X}$ が近づくように, 以下の損失関数 L を最小化する．第 1 項がヒンジ損失，第 2 項が二乗誤差である．

$$\begin{aligned} L &= \max\{0, 1 - \sum_{i,k} C_{i,k} \bar{X}_{i,k} + \sum_{i,k} C_{i,k} X'_{i,k}\} \\ &\quad + \sum_{i,k} (X_{i,k} - \bar{X}_{i,k})^2, \\ X'_{i,k} &= \begin{cases} X_{i,k} & (\max_{i'} X_{i',k} = X_{i,k}) \\ 0 & (\text{otherwise}). \end{cases} \end{aligned} \tag{5}$$

5 実験

“*Urban Animals*”カテゴリ[3]に属し、かつ、2 枚以上の画像を含む英語 Wikipedia 記事からデータを構築し，実験を行った．表 1 がデータセットの概要である．

表 1: 使用データ概要． 画像， セクション数は 1 記事あたりの平均値．

Urban Animals	記事数	画像数	セクション数
training	101	9.4	16.7
validation	34	14.3	22.3

5.1 モデル

文章エンコーダ E_s には，Skip-Thought[14][4]，または，Trigram-Word[13][5]の学習済みモデルを使用した．両エンコーダ共，入力には文を想定するため，セクションは全ての文を繋げ，全体で 1 文として入力する．v_{s_i} の次元 l は，Skip-Thought の場合 2400， Trigram-Word の場合 600 となる．画像エンコーダ E_p には，ResNet34 を使用し，v_{p_k} の次元 r は 512 とした．MLP は隠れ層を 2 層 200 次元とし，活性化関数には ReLU を用い，出力層を 50 次元とした．最適化法は Adam を採用している．

[3] https://en.wikipedia.org/wiki/Category:Urban_animals

[4] https://github.com/tensorflow/models/tree/master/research/skip_thoughts

[5] https://github.com/jwieting/para-nmt-50m

5.2 結果

表 2: 実験結果

	画像 + キャプション		画像	
	ST	TW	ST	TW
Top 1	0.161	0.171	0.140	0.180
Top 3	0.365	0.375	0.344	0.398
Top 5	0.502	0.524	0.506	0.540

全画像の内， 類似度上位 1・3・5 位に正解セクションが入っていたものの割合．
ST : Skip-Thought，TW : Trigram-Word

実験結果を表 2 に示す．キャプションの有無によらず， 文章エンコーダとして Trigram-Word の方が Skip-Thought よりも正しく挿入セクションを推定できている．ただし，Trigram-Word は画像情報としてキャプションを利用すると精度が下がり（Top 1 : 0.180 → 0.171），画像ベクトル v_p とキャプションベクトル v_p^{cap} を合わせた情報を上手く活用できていない．一方で，Skip-Thought ではキャプション情報によって精度が向上しており（Top 1 : 0.140 → 0.161），単純にベクトルの結合が不十分というわけではなく，手法との組み合わせにも依存している．

5.3 考察

本節では Trigram-Word を用いた際の，実際の挿入成功例・失敗例を分析し，現時点での困難点や今後の課題について考察する．

図 2 が成功例である．左がシマスカンクのつがいと巣の画像であり、シマスカンクの繁殖や子育てに関するセクション（*Reproduction and development*）に挿入できている．右は鹿狩りを描いた浮世絵であり，狩りに関するセクション（*Hunting*）に挿入できている．両者とも画像のみで正しく推定できており，これは，「つがい」「巣」の認識から「繁殖」「子育て」といった推論，画像の細部から「弓矢」や「倒れる鹿」の意味抽出といったモデルの高度な理解を示唆している．

一方，図 3 が失敗例であり，オポッサムの特徴的な行動‘*play possum*（死んだふり）’を写したものである．こちらの画像のみから「死んだふり」を汲み取るのは難しい．しかし，キャプションを用いて，正解セクションとの間で共通する特徴的な語（e.g. *play possum, fake death, injured*）を利用できるようにしても，正解できなかった．これは異なるメディアからの情報を取捨選択しながら正解を導くことの難しさを示唆している．

図 2: 成功例
キャプション／正解セクション見出し（左）: *Striped Skunk Pair ／ Reproduction and development*
キャプション／正解セクション見出し（右）: *Tsukioka Yoshitoshi Ukiyo-e depicting the Minamoto no Tsunemoto hunting a sika with a yumi ／ Hunting*

図 3: 失敗例
キャプション／正解セクション見出し：*When injured or threatened, the Virginia opossum is well known for attempting to fake death or "play possum", as seen in this photo. ／ Behavior*

6 結論

本稿では，記事に対する画像挿入という新たなタスクを提案し，ベースラインとなる手法を考案，性能評価を行った．提案タスクでは，セクションや画像が様々に関連し合うこと，そして，その関連性から全体としての意味を捉えることが重要であり，その点が既存のタスクとは異なる特徴である．マルチモーダルな環境において，このようにモデルの深い理解を問う課題は，今後の機械による意味理解という目標に対して，重要な出発点となるだろう．

参考文献

[1] Harsh Agrawal, Arjun Chandrasekaran, et al. Sort story: Sorting jumbled images and captions into stories. In *EMNLP*, 2016.

[2] Antoine Bosselut, Jianfu Chen, et al. Learning prototypical event structure from photo albums. In *ACL*, 2016.

[3] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014.

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017.

[5] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, Vol. 16, No. 12, pp. 2639–2664, 2004.

[6] Y. He, S. Xiang, et al. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, Vol. 18, No. 7, pp. 1363–1377, 2016.

[7] Mohit Iyyer, Varun Manjunatha, Anupam Guha, et al. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *CVPR*, 2017.

[8] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *PMLR*, 2014.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[10] Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. Deeper lstm and normalized cnn visual question answering model. `https://github.com/VT-vision-lab/VQA_LSTM_CNN`, 2015.

[11] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 9, pp. 2372–2385, 2018.

[12] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pp. 3156–3164, 2015.

[13] J. Wieting and K. Gimpel. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *ACL*, 2018.

[14] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.

[15] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 67–78, 2014.