属性付き画像キャプション生成

理学専攻・情報科学コース
2040663
孫　延君

1　はじめに

人工知能における実世界理解を促進するためには，人間のようにマルチモーダル情報を解釈し，推論する研究が必要である [1]．また，単一モダリティに着目しても，自然言語は抽象的な概念を表現できる性質を有し，他のモダリティ情報を表現することができることから，他のモダリティに関連づけた情報処理に多用される．このような背景から，本研究では，画像を用いた文の復元や同義文への変換を行い，マルチモダリティは単一モダリティよりも復元や変換の精度が向上られることを検証した．加えて，画像とキャプション両方の情報から画像キャプション生成を行った．

2　画像を通じた同義文の潜在空間における対応関係

自然言語では，主語と目的語を入れ替えることで表現は異なるが同義になる文が存在する．また，これらの文が指し示す画像は同じものとなり，画像が表す意味とそれぞれの文が表す意味との間には何らかしらの共通性が存在すると考えられる．本研究は，その共通性を用い，同じ意味を持つ文が共通に指し示す画像を軸として，それらの変換を行ったことにより，単一モダリティよりもマルチモダリティの情報を使った方が自然言語理解の精度が向上られることを検証した．

2.1　問題設定

画像情報が自然言語理解に有用であるかを考察するため,「画像なし」と「画像あり」で入力文の復元および同義文へ変更することを行った．2 つの数字と数字間の位置関係を記述する同義文 2 文 $\boldsymbol{X}$，$\boldsymbol{Y}$ と双方の文が指し示す内容を有する共通の画像 $\boldsymbol{I}$ をペアデータ $\mathcal{D}$ とする．その内訳を以下に示す．

$$\mathcal{D} = \{\{\boldsymbol{X_1}, \boldsymbol{Y_1}, \boldsymbol{I_1}\}, ..., \{\boldsymbol{X_n}, \boldsymbol{Y_n}, \boldsymbol{I_n}\}\} \quad (1)$$

$$\boldsymbol{X_n} = [X_n^{(1)}, X_n^{(2)}, ..., X_n^{(t)}] \quad (2)$$

$$\boldsymbol{Y_n} = [Y_n^{(1)}, Y_n^{(2)}, ..., Y_n^{(t')}] \quad (3)$$

$$\boldsymbol{I_n} = [I_n^{(1)}, I_n^{(2)}, ..., I_n^{(d)}] \quad (4)$$

ここで，　$n, t, t', d \in \mathbb{N}$．$n$ はデータの数，t は文の長さ，d は画像のピクセル数，$\mathbb{N}$ は自然数である．

「画像なし」の設定は，文 $\boldsymbol{X}$ の情報のみを用いて同じ文の復元や同義文の変換を行う.「画像あり」の設定は，画像を含め $\boldsymbol{X}, \boldsymbol{I}$ から $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I}$ への復元および同義文への変換を行う．

2.2　モデル設計

画像モダリティの情報を処理するため，変分オートエンコーダ（VAE）モデルを基づい，Encoder と Decoder を畳込みニューラルネットワーク（CNN）に置き換え，VAE-CNN を構築した．また，自然言語モダリティの情報処理は HR-VAE[2] を採用した． HR-VAE と VAE-CNN の Encoder によって自然言語モダリティと画像のモダリティの潜在変数 z_X と z_I を得る．潜在変数 z_X，z_I はそれぞれのモダリティ潜在空間に正規分布に従う．続いて，z_X，z_I を用い，共通潜在変数 z_c を作成した．z_c の作り方が 2 つある．一つ目は，z_X，z_I を合併した手法，concat という．二つ目は，PoE[3] を用いた，複数モダリティの情報を一つモダリティと見なし，z_X, z_I 潜在変数の分散や平均をそれぞれ合併し，新しい正規分布を作成する手法である．ここに，その手法は joint と言う．最後に，VAE-CNN，HR-VAE の Decoder を用い，共通潜在変数 z_c に沿って画像とテキストの復元および同義文への変換を行った．

2.3　実験

本研究は，MINST データセット[1]を使って，実験を行った．

評価指標として，METEOR と正解率を用いる．正解率はテキスト復元と同義文への変換した際の完全復元と変換の数によって評価する．

実験結果は表 1 と表 2 に示す．結果によって，画像を通じたテキストの再構成と同義文の変換は，テキストのみからなる同義文変換よりも効率の良い学習と，精度の高いテキストの完全復元及び変換を実現した．

表 1: テキストの復元及び同義文への変換の METEOR

Num. of Training Data	10,000				30,000				60,000			
Experiment	Reconstruct		Transform		Reconstruct		Transform		Reconstruct		Transform	
Epoch	200	500	200	500	200	500	200	500	200	500	200	500
HR-VAE	75.67	75.93	64.70	77.34	69.35	69.39	70.45	69.27	76.65	76.07	74.30	75.82
+(concat) VAE-CNN	**93.92**	**96.50**	**93.92**	**96.48**	**96.10**	**95.04**	**95.94**	**95.05**	**96.74**	**97.01**	**96.81**	**97.13**
+(joint) VAE-CNN	88.50	83.72	85.00	85.92	86.62	89.13	85.63	86.23	87.29	86.91	85.43	84.58

表 2: テキストの復元及び同義文への変換の正解率（%）

Num. of Training Data	10,000				30,000				60,000			
Experiment	Reconstruct		Transform		Reconstruct		Transform		Reconstruct		Transform	
Epoch	200	500	200	500	200	500	200	500	200	500	200	500
HR-VAE	0	0	0	0	0	0	0	0	0	0	0	0
+(concat) VAE-CNN	**31.45**	**38.75**	**30.23**	**38.73**	**46.59**	**50.11**	**47.48**	**50.31**	**54.53**	**56.63**	**53.72**	**57.01**
+(joint) VAE-CNN	9.425	20.1	6.475	8.575	19.62	24.16	12.31	18.22	32.97	14.68	22.90	16.42

3　同義文からの画像生成

2.3 節の実験によって，潜在空間におけるこれらの同義文と対象画像は同じように扱われると想定する．同義文をそれぞれ用い，画像の復元を行ったが，この共通潜在変数に画像情報も含んだ．そして，同義文による類似な画像生成できるに関するは有効な証拠と言えないと考えられる．本節では，同義文による類似画像を生成できることをさらに証明するため，第 2 章と同じペアデータを用い，同義文からの画像生成の実験を行なった．

具体的には，まずはテキスト X により画像 I' を生成し，続いて，画像 I' による同義文 Y' 生成し，最後に生成された同義文 Y' を用いて画像 I'' 生成を行う．これて，画像を入力なしでの設定の上，同義文 2 つか

[1] http://yann.lecun.com/exdb/mnist/

らそれぞれ画像を生成する．また，生成された画像 I' と I'' の類似度によって同義文の対応関係を考察する．

3.1 実験

本研究では，同じく，MNIST を使って実験を行った．

評価指標では，同一な画像を軸とし，テキスト X から同義である文 Y への変換の結果を考察するため，BLEU，METEOR と正解率を使って評価する．また，同義である文から生成された画像の cos 類似度を測る．

実験結果は表 3 に示す．結果によって，画像を軸とし，同義文間の変換が可能であることを証明した．また，同義文が同じ意味を表現するため，潜在空間において，同義文が同じように扱われていることをわかった．

表 3: 各評価指標による同義文からの画像生成の結果の評価

Num. of Training Data	BLEU	METEOR	Correct Rate	Cosine Similarity
2,000	0.9914	0.9851	0.8181	0.9460
6,000	0.9894	0.9825	0.7981	0.9794
10,000	0.9863	0.9775	0.7250	0.9735

4 属性付き画像キャプション

4.1 画像キャプション

近年，画像中に含まれる人や物などの物体とその属性，及び，物体間の関係に注目したキャプション生成を目指し，シーングラフを用いたキャプション生成の研究は多く行われている [4, 5]．これらの研究では，画像内に写っている内容を正確に記述可能になったが，人は画像を見る際に，見た目の形容詞 (例：可愛い，美味しい) を無意識にしゃべる．本研究は，画像とキャプションからのシーングラフを統合し，もっと人が表現するような形容詞をつけた豊富な画像キャプションの生成を行う．

4.2 シーングラフを用いたキャプション生成

シーングラフは，画像中のオブジェクト，関係，属性をノードとし，それらの間の関係をを有向辺として表現した有向グラフである．シーングラフの構築に先立ち，まずは画像内に含まれる物体の認識が行われる．本研究は Faster-RCNN を用い，物体認識を行う．また，物体の属性を認識するため，RoI-Pooling をもう 1 つを設けた．次に，Neural-Motifs は Faster-RCNN に基づいた画像からのシーングラフ生成器である．本研究は，Neural-Motifs の Faster-RCNN を上記に説明したように改良し，画像からの属性付きシーングラフ S^I を作成する．同時に，Stanford Scene Graph Parser[2]を用いて，キャプションからのシーングラフ S^S を作成する．最後に，画像シーングラフからの物体ノードを軸とし，属性ノードとキャプションシーングラフの関係ノードを統合し，キャプション生成用のシーングラフ S^C になる．最後に，S^C を LSTM に入力し，キャプション生成を行う．
提案手法は図 1 に示した通りである．

4.3 実験

画像シーングラフを生成することをトレーニングするため，Visual Genome データセットを使った．画像キャプション生成のトレーニングに対しては，MS COCO

[2]https://nlp.stanford.edu/software/scenegraph-parser.shtml

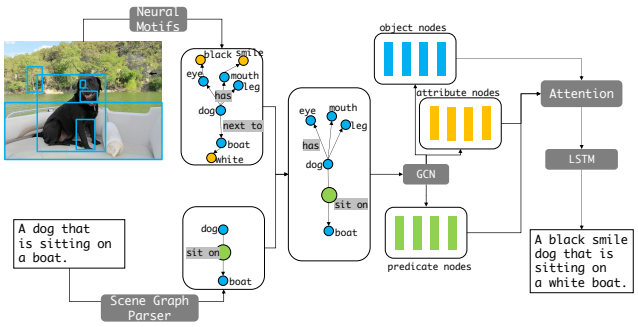

図 1: 提案する画像キャプションのフレームワークの概要図

表 4: 各評価指標による生成したキャプションの評価結果

Models	BLEU@1	BLEU@4	METEOR	ROUGE_L	CIDEr	SPICE
Weakly-VRD	77.06	35.55	27.71	56.77	110.98	21.06
Ours	**77.67**	**36.07**	**27.78**	**56.96**	**112.28**	**21.20**

と COCO Attribute データセットを使った．訓練データ 72,786，テストデータ 3,219 である．

属性認識の精度とキャプションの質を評価するため，mAP と画像キャプションの先行研究によく使われている BLEU, METEOR, ROUGE_L, CIDEr, SPICE を用いて評価した．

4.4 結果

属性認識の mAP は 0.56 だった．COCO Attribute の元論文に提案した属性認識の精度 0.14 より多く向上した．画像キャプションの結果は表 4 に示す．また，シーングラフを用いた画像キャプション生成モデル Weakly-VRG[5] を比較モデルとする．表 4 によって，本研究提案したモデルは比較モデルより質が良いキャプション生成ができた．

5 おわりに

本研究では，同じ画像を軸として，文の復元や同義文への変換により，マルチモダリティを用いた自然言語理解の精度が良いと検証した．また，同義文が同じ意味を表現するため，潜在空間において，同義文が同じように扱われていることをわかった．加えて，画像とキャプションのシーングラフから統合したシーングラフを用いて，属性付き画像キャプションの生成を行った．

参考文献

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp. 423–443, 2019.

[2] Ruizhe Li, Xiao Li, Chenghua Lin, Matthew Collinson, and Rui Mao. A stable variational autoencoder for text modelling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 594–599, Tokyo, Japan, October–November 2019. Association for Computational Linguistics.

[3] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, Vol. 14, No. 8, p. 1771–1800, August 2002.

[4] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10677–10686, 2019.

[5] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7454–7464, 2020.