ゲート付き相互注意を用いたエンコーダ・デコーダによる感情に基づく絵画説明文生成

石川慎太朗　杉浦孔明
慶應義塾大学
{shin.0116,komei.sugiura}@keio.jp

概要

絵画作品鑑賞のバリアフリー化を進める際に，人手を必要とせずに絵画の説明文を生成できれば有用である．絵画の解釈は鑑賞者の感情と密接な関わりを有するため，説明文は感情に基づいていることが望ましい．本研究では，感情ラベルを視覚情報に統合する Affective Visual Encoder を導入した絵画説明文生成モデルを提案する．提案手法においては，感情トークンを用いて画像の領域・グリッド特徴量を融合し，画像・物体レベルの視覚情報を利用する．ArtEmis データセットを使用して提案手法の性能を評価した結果，全ての評価指標において既存手法を上回る性能を得た．

1　はじめに

絵画鑑賞においては，視覚障害者へのバリアフリー化が長年にわたって推進されており [1]，多くの絵画には歴史等の事実に基づいた説明文が付与されている．一方，絵画の解釈と鑑賞者の感情との間には密接な関係性が存在すると指摘されているため [2]，感情が反映された説明文は絵画鑑賞において重要である．しかし，全ての絵画について感情を考慮した説明文が与えられているわけではない．

そこで本研究では，絵画に対する鑑賞者の感情を考慮した説明文を生成するモデルを構築する．図 1 に例を示す．図中の絵画の画像および “excitement” という感情が与えられたときに，鑑賞者の感情の高ぶりを反映させた “The ships look like they are about to go on an adventure.” という説明文を生成することが期待される．

多くの既存研究 [3, 4] は，利用者が指定した感情に基づく説明文の生成には取り組んでおらず，画像に関連の深い感情を予測した後，それを追加入力として説明文の生成に利用するというアプローチを採用している．しかし，これらの手法においては任意の感情を入力として受け取ることはできず，感情で条件付けされた説明文を生成することはできない．

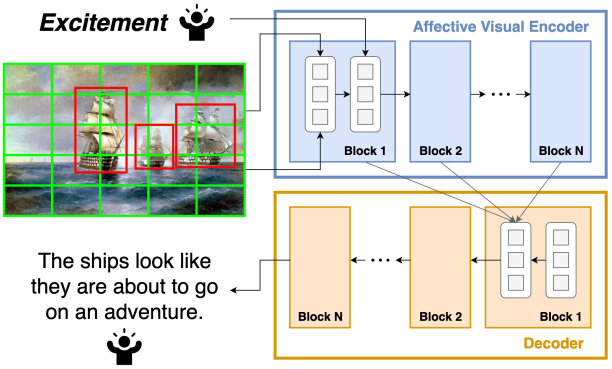

図 1　提案手法の概要

本研究では，Affective Visual Encoder (AVE) を導入した絵画説明文生成モデルを提案する．提案手法では，CLIP[5] を使用して画像から抽出したクロスモーダル情報を入力に追加することで，視覚情報を強化する．以上により，絵画の内容と感情の種類を適切に反映させた説明文を生成することが期待される．図 1 に提案手法の概要を示す．既存手法と異なる点は，AVE を用いて，感情ラベルを視覚情報に統合する点である．

本研究の貢献を以下に示す．

- ゲート付き相互注意機構を用いて，視覚情報を感情ラベルで条件付けする AVE モジュールを提案する．
- エンコーダにおいて，画像の領域・グリッド特徴量を融合するための感情トークンを導入する．

2　問題設定

本研究で扱うタスクを，Affective Image Captioning (AIC) と定義する．AIC タスクは，絵画と感情が与えられたときに，絵画の内容および感情の種類に

沿った説明文を生成するタスクである．

当該タスクの入出力を以下のように定義する．

- **入力:** 絵画の画像，感情ラベル
- **出力:** 感情を反映させた絵画の説明文

本論文で使用する用語を以下のように定義する．

- **感情:** ArtEmis データセットにおいて定義されている感情ラベル群のうちの 1 つ．感情ラベルは，“anger” “disgust” “fear” “sadness” “amusement” “awe” “contentment” “excitement” “something else” の計 9 種類存在する．

モデルの評価には，標準的な自動評価尺度である BLEU[6]，METEOR[7]，ROUGE-L[8]，CIDEr[9]，SPICE[10] を使用する．

3 提案手法

提案手法のネットワーク構造を図 2 に示す．ネットワークは複数のエンコーダ・デコーダブロックから構成される．各モジュールの詳細については本節で後述する．

本手法は，Meshed-Memory Transformer (M2)[3] 等の既存の説明文生成手法と関連が深い．M2 は，QKV 注意機構を使用したエンコーダ・デコーダ型の構造を有しており，ArtEmis データセット [11] において高い性能が報告されている．本手法は，M2 をもとにしているものの，GRIT[4] 等の他の説明文生成手法にも適用可能である．

3.1 入力

ネットワークの入力 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = \{X_{\text{grd}}, X_{\text{rgn}}, X_{\text{emo}}, X_{\text{jnt}}\}, \tag{1}$$

$$X_{\text{grd}} = \{X_{\text{grd}}^{(i)} | i = 1, 2, ..., N\} \tag{2}$$

ここに，X_{grd} はグリッド特徴量，$X_{\text{rgn}} \in \mathbb{R}^{N_{\text{rgn}} \times d_{\text{rgn}}}$ は領域特徴量，$X_{\text{emo}} \in \mathbb{R}^{d_{\text{emo}}}$ は感情の埋め込み表現，$X_{\text{jnt}} \in \mathbb{R}^{d_{\text{jnt}}}$ は画像のクロスモーダル表現を表す．なお，$X_{\text{grd}}^{(i)} \in \mathbb{R}^{M \times d_{\text{grd}}^{(i)}}$ はバックボーンネットワークから得られる i 番目の特徴マップ $\boldsymbol{s}_t$ を表し，N はエンコーダのブロック数を表す．

本手法では，事前学習済みの Swin Transformer[12] と CLIP[5] を使用して，それぞれ画像の特徴マップとクロスモーダル表現を獲得する．また，Visual Genome データセット [13] で学習済みの Faster R-CNN[14] を使用して，各画像から領域特徴量を抽出する．さらに，感情ラベル E を全結合層に入力することで，埋め込み表現を獲得する．

3.2 Affective Visual Encoder

AVE は複数のブロックから構成される．当該モジュールは，モデルへの入力をもとに，感情トークンを使用して画像・感情特徴量を融合する．

各エンコーダブロックにおいては，画像と感情の関係性をモデル化するために相互注意機構を使用する．

相互注意と自己注意の処理を，行列 X_{A} および X_{B} を用いて以下のように定義する．

$$\text{CrossAttn}(X_{\text{A}}, X_{\text{B}}) = \text{softmax}(\frac{(W_{\text{Q}}X_{\text{A}})(W_{\text{K}}X_{\text{B}})^{\top}}{\sqrt{d}})(W_{\text{V}}X_{\text{B}}), \tag{3}$$

$$\text{SelfAttn}(X_{\text{A}}) = \text{CrossAttn}(X_{\text{A}}, X_{\text{A}}) \tag{4}$$

ここに，$W_{\text{Q}} \in \mathbb{R}^{d_{\text{q}} \times d_{\text{in}}}$，$W_{\text{K}} \in \mathbb{R}^{d_{\text{k}} \times d_{\text{in}}}$，$W_{\text{V}} \in \mathbb{R}^{d_{\text{v}} \times d_{\text{in}}}$ はいずれも学習可能な重みである．また，次のエンコーダブロックに入力される情報量を制御するために，以下のようなゲート構造を導入する．

$$f_{\text{gate}}(X) = X \odot \tanh(W\boldsymbol{h}) \tag{5}$$

ここに，$\boldsymbol{h}$ は中間層の出力，W は学習可能な重みであり，$\odot$ はアダマール積を表す．

i 番目のエンコーダブロックでは，はじめに $\boldsymbol{h}_{\text{tok}}^{(i)} = \text{SelfAttn}(Z^{(i)})$ という処理を実行する．$Z^{(1)} = Z$ は感情トークンを表す．次に，グリッド特徴量 $X_{\text{grd}}^{(i)}$ をもとに $\boldsymbol{h}_{\text{tok}}^{(i)}$ にゲート付き相互注意演算を適用する．

$$\boldsymbol{h}_{\text{grd}}^{(i)} = f_{\text{grd\_gate}}(\text{CrossAttn}(\boldsymbol{h}_{\text{tok}}^{(i)}, W_{\text{grd}}^{(i)}X_{\text{grd}}^{(i)})) \tag{6}$$

ここに，$\{W_{\text{grd}}^{(i)} \in \mathbb{R}^{d_{\text{enc}} \times d_{\text{grd}}^{(i)}} | i = 1, 2, ..., N\}$ は学習可能な重みであり，$f_{\text{grd\_gate}}(\cdot)$ は $f_{\text{gate}}(\cdot)$ と同様の構造を有する．また，同様の処理により，X_{rgn} をもとに $\boldsymbol{h}_{\text{rgn}}^{(i)}$ を獲得する．最後に，$\tilde{\boldsymbol{h}}_{\text{tok}}^{(i)} = \text{FFN}(\boldsymbol{h}_{\text{grd}}^{(i)} + \boldsymbol{h}_{\text{rgn}}^{(i)})$ という処理を実行する．

X_{emo} および X_{jnt} は，条件付けモジュールによって視覚情報に統合される．はじめに，相互注意機構により X_{emo} をエンコードした後，以下のように順伝播型ニューラルネットワークに入力する．

$$\boldsymbol{h}_{\text{emo}}^{(i)} = \text{FFN}(X_{\text{emo}}) \odot \text{FFN}(\text{CrossAttn}(X_{\text{emo}}, \tilde{\boldsymbol{h}}_{\text{tok}}^{(i)})) \tag{7}$$

続いて，$\tilde{\boldsymbol{h}}_{\text{emo}}^{(i)} = f_{\text{emo\_gate}}(\boldsymbol{h}_{\text{emo}}^{(i)})$ という処理を実行する．$f_{\text{emo\_gate}}(\cdot)$ は $f_{\text{gate}}(\cdot)$ と同様の構造を有する．また，同様の処理により，X_{jnt} をもとに $\tilde{\boldsymbol{h}}_{\text{jnt}}^{(i)}$ を獲得する．最後に，$\tilde{Z}^{(i)} = \tilde{\boldsymbol{h}}_{\text{tok}} + \tilde{\boldsymbol{h}}_{\text{jnt}}^{(i)} + \tilde{\boldsymbol{h}}_{\text{emo}}^{(i)}$ という処理によって出力 $\tilde{Z}^{(i)}$ を獲得する．

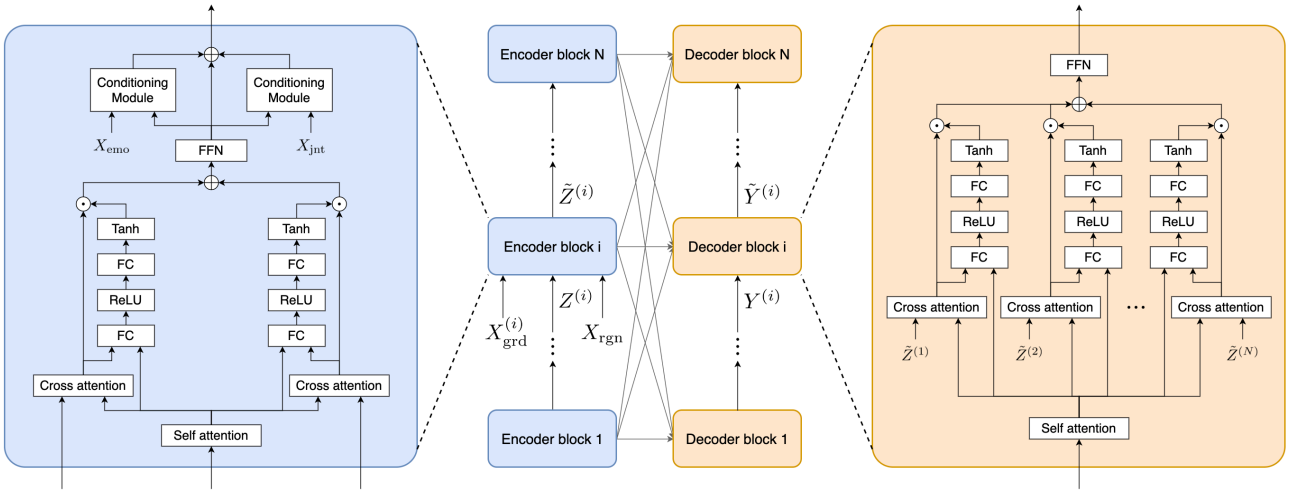

図 2　提案手法のネットワーク構造

3.3　Decoder

エンコーダと同様に，デコーダも複数のブロックから構成される．[3] に従って，各デコーダブロックは，最後のエンコーダブロックの出力のみならず，全てのエンコーダブロックの出力を考慮した相互注意演算を実行する．エンコーダの出力群 $\{\tilde{Z}^{(i)}|i = 1, 2, ..., N\}$ をもとに，デコーダは説明文を生成する．

i 番目のデコーダブロックでは，はじめに前のブロックの出力 $Y^{(i)}$ をマスク付き自己注意機構 [15] に入力した後，$\tilde{Z}^{(j)}$ との間で相互注意演算を実行することで，$\boldsymbol{h}^{(i,j)}_{\text{seq}}$ を獲得する．なお，$Y^{(1)} = \hat{Y}$ は生成された単語群を表し，学習中は Teacher Forcing を採用する．次に，各レベルの出力をゲート構造に入力した後，それらを足し合わせる．

$$\tilde{\boldsymbol{h}}^{(i)}_{\text{seq}} = \sum_{j=1}^{N} f^{(i)}_{\text{seq\_gate}}(\boldsymbol{h}^{(i,j)}_{\text{seq}}) \tag{8}$$

$\{f^{(i)}_{\text{seq\_gate}}(\cdot)|i = 1, 2, ..., N\}$ は $f_{\text{gate}}(\cdot)$ と同様の構造を有する．最後に，$\tilde{Y}^{(i)} = \text{FFN}(\tilde{\boldsymbol{h}}^{(i)}_{\text{seq}})$ という処理によって出力 $\tilde{Y}^{(i)}$ を獲得する．なお，損失関数には交差エントロピー誤差を使用する．

4　実験

4.1　実験設定

本実験では，AIC タスクにおける提案手法の性能を調査した．モデルの評価には ArtEmis データセット [11] を使用した．ArtEmis データセットは，WikiArt から収集した 80,381 枚の絵画および 454,684 の感情と説明文のペアから構成される．語彙サイズは 37,250，全単語数は 7,229,476，平均文長は 15.9 である．本実験では，[11] に従って当該データセットを分割した．訓練集合，検証集合，テスト集合のサンプル数はそれぞれ 338,777，19,931，39,850 である．訓練集合と検証集合を用いてそれぞれモデルのパラメータの更新とハイパーパラメータのチューニングを行い，テスト集合によってモデルの性能を測定した．

本手法で使用したエンコーダ・デコーダは，層数が 3，隠れ層の次元数が 512，Attention Head の数が 8 であり，感情トークンの数 N_{tok} は 50 であった．最適化には Adam を使用し，エポック数は 20，バッチサイズは 32 であった．

提案モデルのパラメータ数は 6800 万であった．学習には，メモリ 32GB 搭載の Tesla V100 を 4 台使用し，12 時間を要した．また，1 つの説明文の生成に要する時間は 0.03 秒であった．学習中は，各エポックにおいて検証集合およびテスト集合によるモデルの評価を行い，検証集合において最も METEOR スコアが高かったときのテスト集合におけるスコアを，最終的なモデルのスコアとした．

4.2　定量的結果

ベースライン手法との比較に関する定量的結果を表 1 に示す．各スコアは，5 回の実験における平均値および標準偏差を表す．本実験では，Show Attend Tell (SAT)[16] および M2[3] をベースライン手法とした．これは，当該手法が代表的な説明文生成手法であり，ArtEmis において良好な結果が報告されているためである．

表 1　ベースライン手法との比較に関する定量的結果

Method	B-4	M	R	C	S
SAT[16]	$3.1_{\pm 0.0}$	$8.9_{\pm 0.0}$	$17.8_{\pm 0.3}$	$12.8_{\pm 0.2}$	$6.6_{\pm 0.2}$
M2[3]	$3.0_{\pm 0.0}$	$8.8_{\pm 0.1}$	$19.0_{\pm 0.2}$	$13.8_{\pm 0.2}$	$7.6_{\pm 0.1}$
Ours	$\mathbf{3.3}_{\pm 0.1}$	$\mathbf{9.2}_{\pm 0.1}$	$\mathbf{19.5}_{\pm 0.1}$	$\mathbf{15.4}_{\pm 0.2}$	$\mathbf{8.3}_{\pm 0.1}$

表 2　Ablation studies に関する定量的結果

Cond.	CM	N_{tok}	B-4	M	R	C	S
full	✓	50	3.3	**9.2**	19.5	15.3	8.2
(i)		50	3.2	9.0	19.4	14.5	7.9
(ii)-a	✓	10	3.2	9.0	19.2	15.0	8.1
(ii)-b	✓	25	**3.4**	9.1	19.4	15.4	8.2
(ii)-c	✓	100	3.3	**9.2**	**19.6**	**15.6**	**8.3**
(ii)-d	✓	200	3.3	9.0	19.4	15.0	8.1

2 節で言及した自動評価尺度群のうち，[17] に従って CIDEr および SPICE を主要評価尺度に定めた．以降は各手法の CIDEr スコアのみに言及する．提案手法は 15.4 ポイントである一方，SAT と M2 はそれぞれ 12.8 ポイントと 13.8 ポイントであった．したがって，提案手法は SAT と M2 をそれぞれ 2.6 ポイントと 1.6 ポイント上回った．

4.3　Ablation Studies

Ablation Studies には，以下の 2 条件を定めた．

(i)W/o conditioning modules: 条件付けモジュールを用いた感情・視覚情報の統合手法の代わりに，各エンコーダブロックの最後でそれぞれの特徴量を単純に連結する場合に対して，性能差を調査した．

(ii)More/fewer affective tokens: N_{tok} を増減させることで，性能への影響を調査した．本実験では，N_{tok} が 10，25，100，200 の場合の性能を比較した．

表 2 に結果を示す．“CM” は “Conditioning Module” を表す．(i) の条件では CIDEr スコアが 14.5 であった．full の条件と比較してスコアが 0.8 ポイント下回っていることから，条件付けモジュールが感情の埋め込み表現と画像特徴量のモデル化に有効であると考えられる．また，(ii)-a，(ii)-b，(ii)-c，(ii)-d の条件では，CIDEr スコアがそれぞれ 15.0，15.4，15.6，15.0 であった．したがって，提案手法は感情トークンが 100 のときに最良のスコアを達成した．

(a)

Fear

Ref.: This guy has a very scary look on his face, like he's a mysterious stranger who is up to no good.

M2: The man in the painting looks like he is angry about something.

Ours: This man looks like he's up to no good since he's wearing a suit.

(b)

Amusement

Ref.: The sway of the trees seems like they are alive. The colors are vibrant and lively.

M2: The tree looks like it's growing out of the ground.

Ours: The trees look like they are dancing in the wind.

図 3　定性的結果

4.4　定性的結果

図 3 に定性的結果を示す．図 3-(a) では，E =“fear” であった．提案手法は “This man looks like he’s up to no good since he’s wearing a suit.” という説明文を生成したのに対し，ベースライン手法は “The man in the painting looks like he is angry about something.” という説明文を生成した．ベースライン手法と異なり，提案手法は指定された感情である “fear” を適切に反映できたと考えられる．

図 3-(b) では，E =“amusement” であった．提案手法の生成文は “The trees look like they are dancing in the wind.” であったが，ベースライン手法は “The tree looks like it’s growing out of the ground.” と生成した．提案手法はベースライン手法に比べて，より鑑賞者の感情を考慮した説明文を生成できたと考えられる．

5　おわりに

本研究では，絵画に対する鑑賞者の感情を考慮した説明文を生成するモデルを構築した．本研究の貢献を以下に示す．

- ゲート付き相互注意機構を用いて，視覚情報を感情ラベルで条件付けする AVE モジュールを提案した．
- エンコーダにおいて，画像の領域・グリッド特徴量を融合するための感情トークンを導入した．
- ArtEmis データセットにおいて，提案手法がベースライン手法を全ての評価尺度で上回る性能を得た．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST CREST，NEDO の助成を受けて実施されたものである．

参考文献

[1] Saki Asakawa, João Guerreiro, Dragan Ahmetovic, et al. The Present and Future of Museum Accessibility for People with Visual Impairments. In **ASSETS**, pp. 382–384, 2018.

[2] Leo Tolstoy and Aylmer Maude. **What Is Art?** Oxford University Press, 1959.

[3] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In **CVPR**, pp. 10578–10587, 2020.

[4] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. GRIT: Faster and Better Image captioning Transformer Using Dual Visual Features. In **ECCV**, 2022.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning Transferable Visual Models From Natural Language Supervision. In **ICML**, pp. 8748–8763, 2021.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, 2002.

[7] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In **ACL**, pp. 65–72, 2005.

[8] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **ACL**, pp. 74–81, 2004.

[9] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In **CVPR**, pp. 4566–4575, 2015.

[10] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In **ECCV**, pp. 382–398, 2016.

[11] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, et al. ArtEmis: Affective Language for Visual Art. In **CVPR**, pp. 11569–11579, 2021.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In **ICCV**, pp. 10012–10022, 2021.

[13] Ranjay Krishna, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. **IJCV**, Vol. 123, No. 1, pp. 32–73, 2017.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In **NeurIPS**, Vol. 28, 2015.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. Attention is all you need. In **NeurIPS**, Vol. 30, 2017.

[16] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In **ICML**, pp. 2048–2057, 2015.

[17] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. **arXiv preprint arXiv:2107.06912**, 2021.

[18] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical Sequence Training for Image Captioning. In **CVPR**, pp. 7008–7024, 2017.

[19] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image Captioning: Transforming Objects into Words. In **NeurIPS**, Vol. 32, 2019.

[20] Luowei Zhou, Hamid Palangi, Lei Zhang, et al. Unified Vision-Language Pre-Training for Image Captioning and VQA. In **AAAI**, Vol. 34, pp. 13041–13049, 2020.

[21] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. Multimodal Attention Branch Network for Perspective-Free Sentence Generation. In **CoRL**, pp. 76–85, 2020.

[22] Tadashi Ogura, et al. Alleviating the Burden of Labeling: Sentence Generation by Attention Branch Encoder-Decoder Network. **IEEE RA-L**, Vol. 5, No. 4, pp. 5945–5952, 2020.

[23] Motonari Kambara, et al. Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions. **IEEE RA-L**, Vol. 6, No. 4, pp. 8371–8378, 2021.

[24] Cesc Chunseong Park, Byeongchang Kim, et al. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In **CVPR**, pp. 895–903, 2017.

[25] Cesc Chunseong Park, et al. Towards Personalized Image Captioning via Multimodal Memory Networks. **IEEE Trans. PAMI**, Vol. 41, No. 4, pp. 999–1012, 2018.

[26] Yuya Miyoshi, et al. Automatic Affective Image Captioning System using Emotion Estimation. **Trans. of Japan Society of Kansei Engineering**, Vol. 18, , 2019.

[27] Yiming Zhang, Min Zhang, Sai Wu, and Junbo Zhao. Towards unifying the label space for aspect-and sentence-based sentiment analysis. In **ACL**, pp. 20–30, 2022.

[28] Maarten De Raedt, Fréderic Godin, Chris Develder, et al. Robustifying sentiment classification by maximally exploiting few counterfactuals. In **EMNLP**, 2022.

[29] Tomoyuki Kajiwara, et al. WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In **NAACL**, pp. 2095–2104, 2021.

A　関連研究

説明文生成の分野においては，これまで多くの研究が行われている [16, 18, 19, 20]．代表的なサーベイ論文としては [17] が挙げられる．

Dense Captioning は，自然言語に基づき，画像を領域単位で密に説明するものである．特に [21, 22, 23] は，画像中の物体に関する把持文を生成するモデルを提案している．Personalized Captioning は，利用者が有する知識や語彙，文体等を考慮した説明文の生成に取り組むものである [24, 25, 26]．

自然言語と感情に関する研究も多く実施されている．DPL[27] は，入力文に対してアスペクトベースの感情分類を行う手法である．[28] は，反事実的サンプルを訓練集合に加えることで，感情分類器の頑健性を向上させる研究である．また，WRIME[29] は感情分析の分野における標準データセットであり，SNS 上の日本語の投稿に対する書き手および読み手の感情から構成される．