Tokushima University

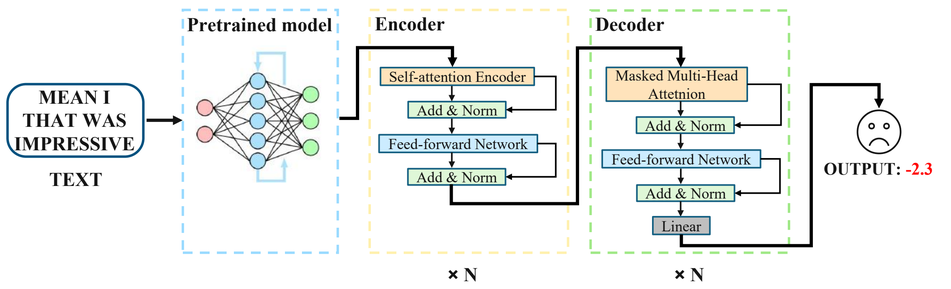
TMAK: Trustworthy Multi-modal Affective AI and Knowledge Engineering Lab

SAGE-Net: Single-layer Augmented Gated Encoder Net work for Efficient Multimodal Sentiment Analysis

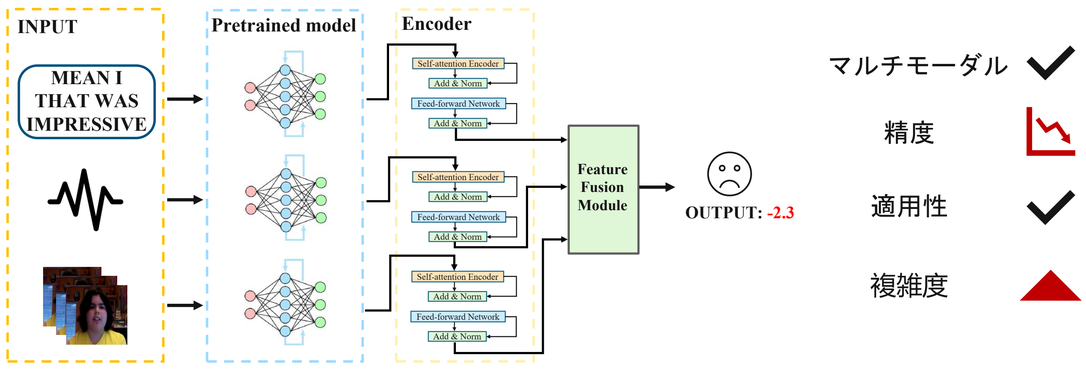
創成科学研究科　創成科学専攻　知能情報・数理科学系プログラム　A2研究室　　周嘉政

研究背景

- 既存の単一モダリティに基づく感情認識手法は、限られた情報に依存するため、実社会における多様で複雑な状況に十分に対応できず、**需要を十分に満たせないだけでなく、精度の面でも依然として不十分である。**



- マルチモーダル手法は複数の情報源を統合することで表現力を高められる利点を有するが、その反面、**計算複雑性が高く、学習や推論に多大なコストを要し、さらに必ずしも高精度を保証するものではない。**



研究目的

本研究では、単一モダリティモデルの精度不足や、マルチモーダルモデルにおける計算コストおよびパラメータ規模の大きさといった課題を解決するため、**軽量化設計**と**効率的な融合戦略**、**データ拡張**を組み合わせた高効率なマルチモーダル感情認識モデルを提案する。

技術的チャレンジ

- 軽量化設計において、**単層エンコーダによる特徴表現の不足をいかに補うか**。
- マルチモーダルフレームワークの設計において、**異なるモダリティ間でいかに効率的かつ効果的に特徴情報の相互作用を実現するか**。
- マルチモーダル特徴融合段階において、**いかに効率的にモダリティ融合を実現するか**。

関連研究

I. マルチモーダル感情分析に関する最新の研究動向

マルチモーダル感情分析の分野では、数多くの高性能モデルにより急速な進展が見られている。たとえば、**MMML（Multimodal Multi-Loss Fusion Network）**、**UniMSE（Unified Framework for Multimodal Sentiment Analysis）**、**CM-BERT（Cross-Modal BERT）**、**SPECTRA（Speech-Text Dialog Pre-training Model）**、および **TEASEL（Speech-Prefixed Language Model）** などのモデルは、それぞれ独自の手法によりマルチモーダル特徴を効果的に捉え、卓越した性能を実現している。

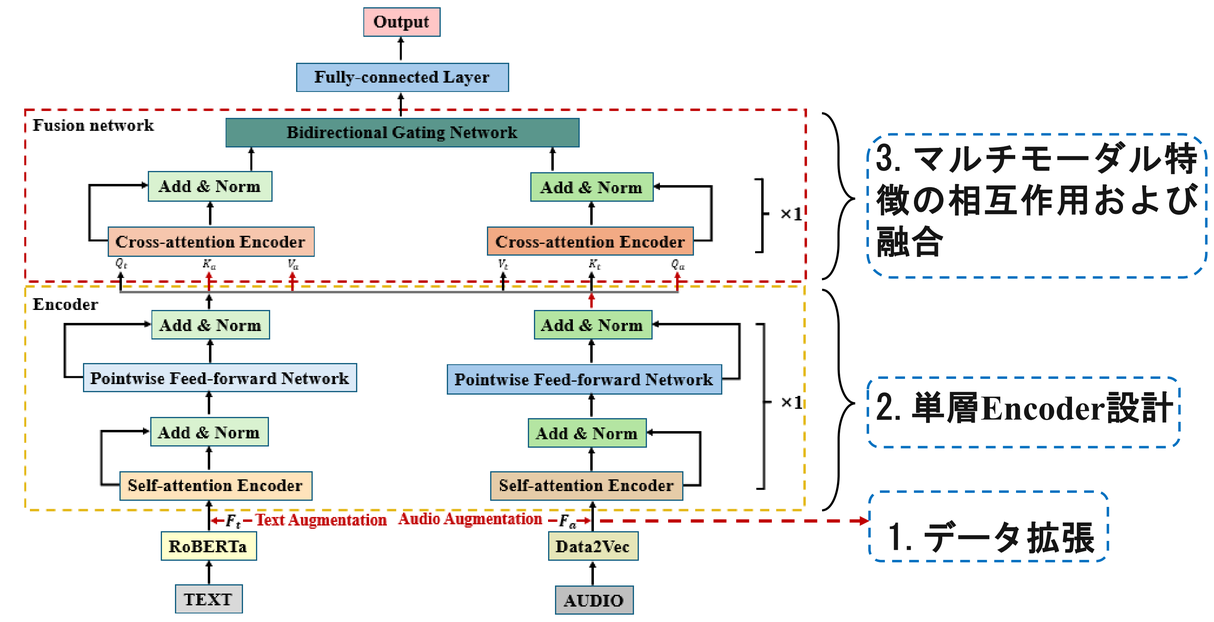
現在のマルチモーダル感情分析モデルの多くは、複数のエンコーダ層を積み重ねるアーキテクチャに依存している。このような構造により、テキスト・音声・映像など各モダリティからの情報を深く抽出・統合でき、感情特徴の理解度が向上する。一方で、こうした多層構造は特徴抽出能力の強化に寄与するものの、**計算コストやパラメータ調整の負担が増大するという**課題も抱えている。

II. 軽量化感情認識モデルに関する研究

軽量化感情認識モデルの研究として、Liuらは**生理信号に基づく感情認識のためのクロスモーダル知識蒸留フレームワーク（EmotionKD）**を提案した。この手法では、脳波（EEG）および皮膚電気反応（GSR）などのマルチモーダル特徴を、単一モダリティであるGSRモデルへと蒸留し、クロスモーダル知識転移を実現する。これにより、マルチモーダルの複雑性を軽減し、リソース制約下での実装が可能となる。一方で、蒸留された単一モダリティのモデルは、他モダリティへの汎化性能が低く、生理信号に特化しているため応用範囲が限定される。

提案手法

I. モデルフレームワークの概要



$$\text{Complexity(Multi-layer)} = O(n^2 \cdot d \cdot L),$$
$$\text{Complexity(Single-layer)} = O(n^2 \cdot d),$$
$$\text{Ratio} = \frac{\text{Complexity(Multi-layer)}}{\text{Complexity(Single-layer)}} = L.$$

本フレームワークは、単層設計により計算複雑性および学習可能パラメータ数を大幅に削減している。

II. データ拡張戦略

Text拡張	同義語置換
ランダムマスキング	ガウス雑音付加
ランダム削除	音量ランダム調整
ランダム挿入	時間マスキング
同義語置換	周波数マスキング

データ拡張の組み合わせ方法を探究することで、モデル性能およびロバスト性の向上を図る。

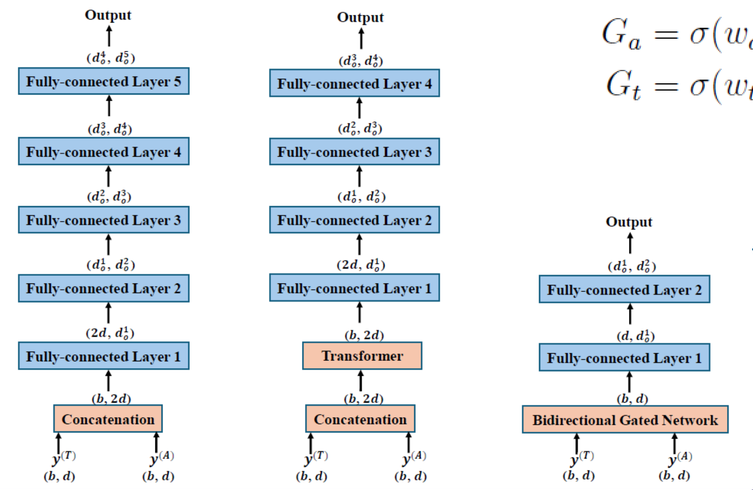
III. 特徴融合

A. 単層クロスアテンション機構により、モダリティ特徴間の効率的な相互作用を実現する。

$$\text{Attention}(Q_t, K_a, V_a) = \text{softmax}\left(\frac{Q_t K_a^\top}{\sqrt{d_k}}\right) V_a,$$

B. 双方向ゲートネットワークにより、特徴の階層的融合を実現する。

$$G_a = \sigma(w_a a + b_a), \qquad W_t = G_a \odot t,$$
$$G_t = \sigma(w_t t + b_t), \qquad W_a = G_t \odot a.$$
$$F = W_t + W_a.$$



本手法は、特徴の効率的な融合を実現するのみならず、**出力特徴の次元を半減させる効果も有している。**

結果と分析

I. SAGE-Net モデル精度

Model	CMU-MOSI					CMU-MOSEI				
	$ACC_{2Has0/Non0}$	$F1_{Has0/Non0}$	ACC_7	MAE	Corr	$ACC_{2Has0/Non0}$	$F1_{Has0/Non0}$	ACC_7	MAE	Corr
COGMEN	-	-/84.34	43.90	-	-	-	-	-	-	-
MMIM	84.14/86.06	84.00/85.98	46.65	0.700	0.800	82.24/85.97	82.66/85.94	54.24	0.526	0.772
TEASEL	84.79/87.50	84.72/85.00	47.52	0.644	0.836	-	-	-	-	-
SPECTRA	-/87.50	-	-	-	-	-/87.34	-	-	-	-
MAG-BERT	84.20/86.10	84.10/86.0	-	0.712	0.796	84.7/-	84.5/-	-	-	-
MISA	81.80/83.40	81.70/83.60	42.30	0.783	0.761	83.60/85.50	83.80/85.30	52.20	0.555	0.756
Self-MM	84.00/85.98	84.42/85.95	-	0.713	0.798	82.81/85.17	82.53/85.30	-	0.530	0.765
UniMSE	85.85/86.90	85.83/86.42	48.68	0.691	0.809	85.86/**87.50**	85.79/**87.46**	54.49	0.523	0.773
MMML	85.91/88.16	85.85/88.15	48.25	0.643	0.838	**86.32**/86.73	**86.23**/86.49	54.95	0.517	0.791
SAGE-Net	**87.03/89.18**	**86.96/89.16**	**50.00**	**0.639**	**0.848**	82.25/86.76	82.82/86.82	**55.27**	**0.516**	**0.796**

Model	ACC_2	ACC_3	ACC_5	F1	MAE	Corr
ALMT	**81.9**	**68.93**	45.73	**81.57**	0.404	0.619
EMT	80.10	67.40	43.50	80.10	0.396	0.623
SAGE-Net	80.09	68.71	**46.17**	80.29	**0.386**	**0.643**

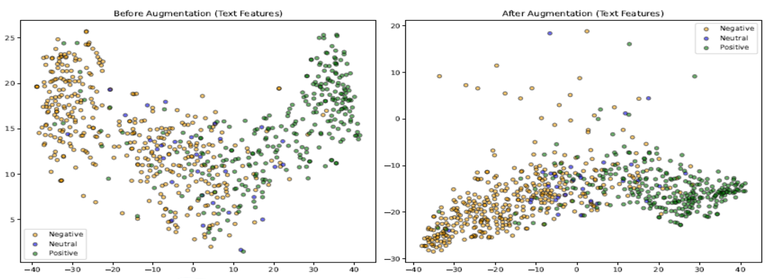
実験は**CMU-MOSEI**、**CMU-MOSI**、および**CH-SIMS**においていずれも良好な結果を得た。

II. メモリ占有量

最大でパラメータを**60.7%削減し**、メモリ使用量を**68.73%低減する**ことができる。

Model	Params (M)	GPU Memory (GB)
EMT	110.5	10.80
MMML	453	11.90
SAGE-Net	178	3.72

III. 特徴分布の状況



特徴拡張はモデル性能を効果的に向上させることができる。