Glorang

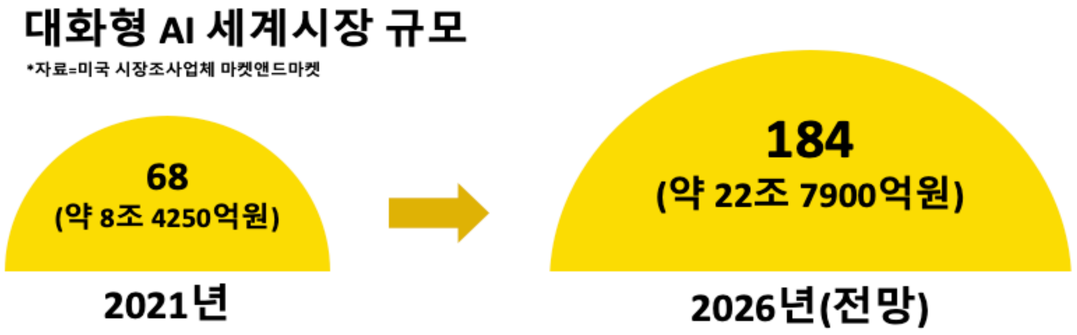
감정 분석을 위한 문맥 및 정렬 기반 음성-문서 퓨전 모델

최성욱 권용훈 서동진
Email: {woody, popul, benny}@glorang.com

Introduction

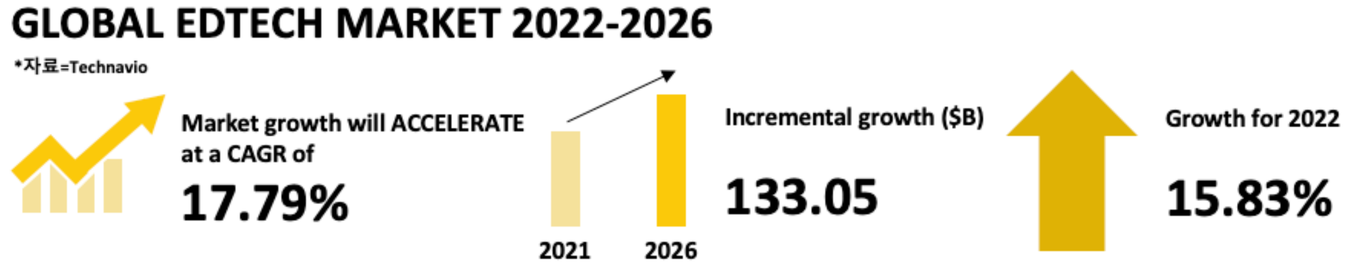
대화형 AI 세계시장 규모

*자료=미국 시장조사업체 마켓앤드마켓



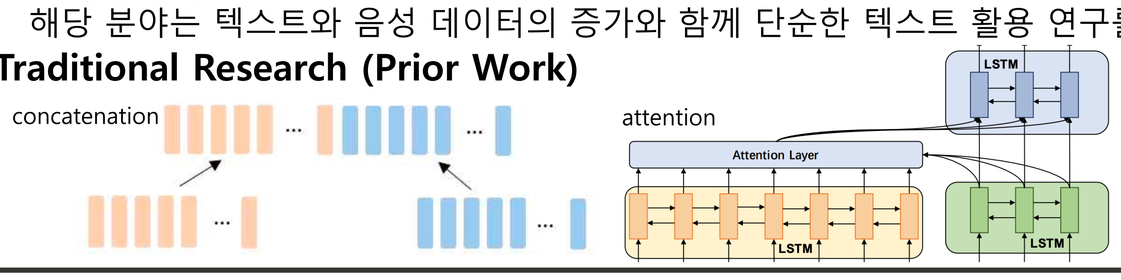
GLOBAL EDTECH MARKET 2022-2026

*자료=Technavio



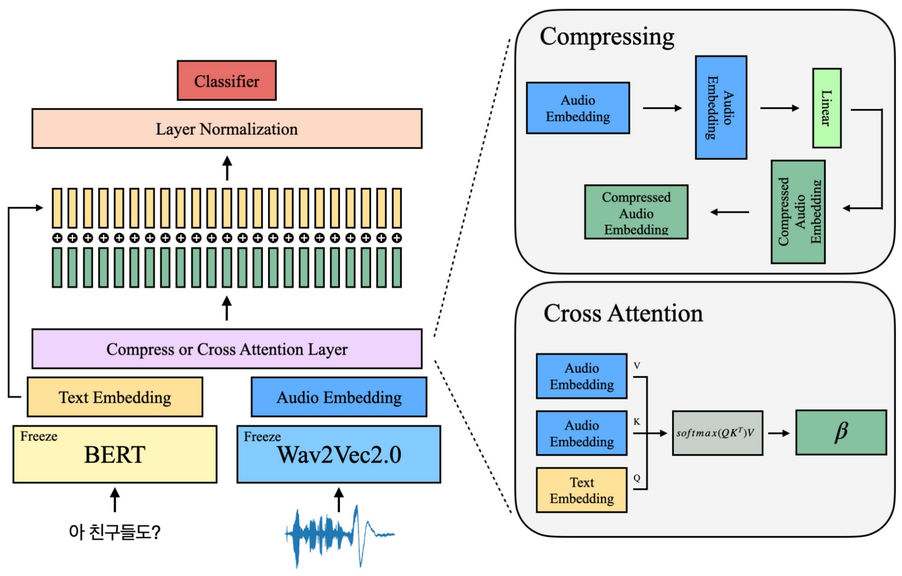
- chatbot을 활용한 온라인 서비스, 화상 통화 플랫폼을 활용한 온라인 교육 시스템의 발전이 급격히 증가하면서 ERC 연구에 대한 관심도 함께 증가하였다.
- ERC 모델은 수업 중 학생들의 감정을 통해 학습 방식에 대한 반응을 파악하고, 이를 교육 서비스의 부가 정보로 사용하여 최적화된 학습 방식을 제공하는 등의 목적으로 활용 가능하다.
- 해당 분야는 텍스트와 음성 데이터의 증가와 함께 단순한 텍스트 활용 연구를 넘어 multi-modal 연구로 확장되고 있다.

Traditional Research (Prior Work)



본 연구에서는 사전 연구와 달리 **서로 다른 모달리티 정보들 사이의 '합'**을 활용하며, 이를 통해 **대응되는 위치의 임베딩 벡터가 서로 정렬되는 방향으로 학습**하도록 한다. 그 결과 concatenation 방식에 비해 성능이 향상되었으며 파라미터 수가 감소하여 높은 효율성을 갖게 되었다.

Method



Compressing

1. Audio Embedding (AE)
2. Transpose AE
3. Pass it to linear layer (Compressed AE)
4. Transpose Compressed AE
5. Adding Compressed AE and Text Embedding

Cross Attention

1. Q(Text), K(Audio), V(Audio) 로써 Dot Attention 연산 적용
2. 생성된 attention weight를 beta값으로 설정하여 Text Embedding과 더함.

Algorithm 1 CASE Training Algorithm

Input : raw text and raw audio in conversation
Output : information of emotion

1: Text → BERT → Text Embedding(TE)
2: Audio → Wav2Vec2.0 → Audio Embedding(AE)
3: Pass them as input into CASE layer
(a) Compressing
- AE(Audio Length x Hidden Dim) → Transpose → AE^T(Hidden Dim x Audio Length)
- AE^T → Linear Layer → Compressed AE^T (Compressing 'audio length' to 'text length')
- Compressed AE^T → Transpose → Compressed AE (Text Length x Hidden Dim)
(b) Attention
- Q: Text Embedding
- K, V: Audio Embedding
- $\beta = softmax(QK^T)V$
4: Adding with Text Embedding

$$\text{Multimodal Embedding} = \begin{cases} TE + \beta, & \text{if } attention \\ TE + CompressedAE, & \text{if } compressing \end{cases}$$

5: Feed multimodal embedding into layernorm layer
6: Pass it into Classifier Head

가정 1

Audio Length를 압축할 때 학습가능한 Layer를 활용하여 Text Length로 압축한다면 **텍스트와 align되는 방향으로 압축**될 것이다.

가정 2

두 모달리티 사이의 정보가 align 되어 있다면 단순히 concatenation을 했을 때보다 addition 했을 때의 성능이 더 좋을 것이다.

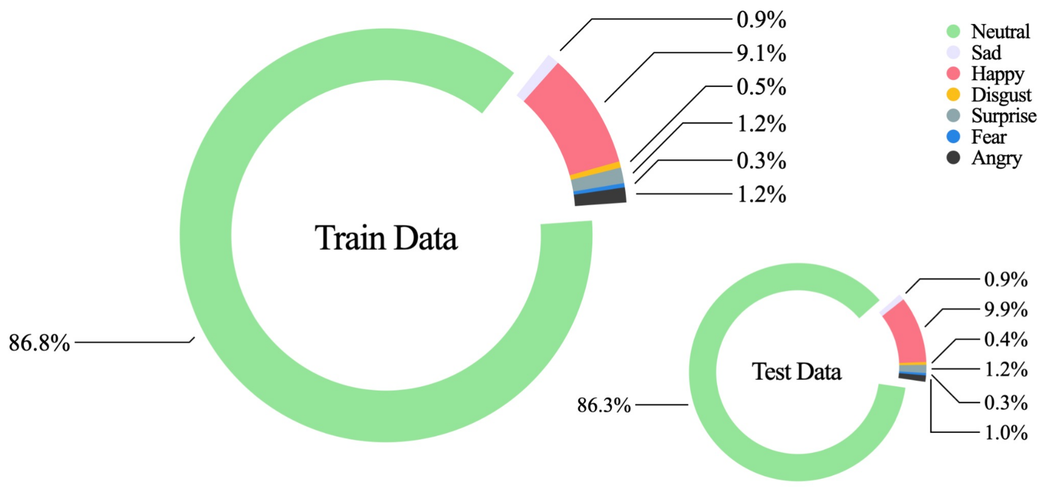
두 가정을 경험적으로 검증하기 위해 다음을 실험한다.

- 길이를 압축하는 학습 가능한 **Compressing Layer**를 활용한 방법을 적용
- Compressing, Cross Attention 방법을 적용하여 **위치별 합 연산의 성능**을 확인

Experiments

Data description

- 텍스트와 음성을 함께 제공하는 Multi-modal 감정 데이터셋
- Label별 비율을 유사하게 학습 및 평가 데이터셋으로 분리
- Neutral label의 비중이 80% 이상인 데이터 불균형이 존재



Metric Settings

- 불균형 데이터셋에 대한 metric으로 Macro-F1 및 Micro-F1 (Neutral label이 있는 데이터를 제외)을 활용

Results

- Compressing을 활용한 CASE Model의 성능이 Macro-F1, Micro-F1 모두에서 가장 좋은 것으로 미루어 보았을 때, **길이를 압축한 방법이 유의미한 성능 향상**으로 이어졌음을 알 수 있다.
- Addition 방법이 Concatenation 방법에 비해 성능이 좋은 것으로 보아, **두 모달리티 사이의 정보가 어느 정도 align 되어있다고 판단**할 수 있다.
- 본 연구에서는 파라미터의 수를 이전 연구에 비해 대폭 줄이면서도 좋은 성능을 달성할 수 있었다.

Models	Macro-F1	Micro-F1	Parameters
Concat	30.16	40.54	
MMM	30.01	40.07	1,547,527
Compressing (Addition)	**32.82**	43.77	**593,031**
Cross Attention (Addition)	**27.91**	**39.51**	**527,367**
Compressing (Concatenation)	31.62	**45.34**	
Cross Attention (Concatenation)	26.49	34.53	

Conclusion

- Multi-modal 감정 인식 모델을 개선하기 위해 서로 다른 modality 정보를 연산하기 위한 위치별 합 방식을 제안함.
- 위치별 합 연산을 통해 서로 다른 modality 사이의 정보를 정렬되도록 구성하는 것이 성능 개선에 유의미한 영향을 끼침.
- Compressing 방법의 성능 향상을 미루어 보았을 때, Audio Embedding을 압축할 때 어느 정도 Text와 align되는 방향으로 압축되는 것을 알 수 있음.
- 이전 연구 대비 Macro-F1에서 약 3%, Micro-F1에서 약 4% 정도의 눈에 띄는 성능 향상 달성.
- 이전 연구 대비 파라미터 수를 대폭 줄이면서 유의미한 성능 향상 달성.

References

[1] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intelligent Systems, vol. 33, no. 6, pp. 17-25, 2018
[2] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," arXiv preprint arXiv:1909.05645, 2019.
[3] D. Krishna and A. Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks.," in Interspeech, pp. 4243–4247, 2020.
[4] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: multi-modal fusion network for emotion recognition in conversation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4652–4661, 2022.
[5] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "Emocaps: Emotion capsule based model for conversational emotion recognition," arXiv preprint arXiv:2203.13504, 2022.
[6] 방나모, 연희연, 이지현, and 구명완, "Mlp-mixer 구조를 활용한 대화에서의 멀티모달 감정 인식," 한국정보과학회 학술발표논문집, pp. 2288-2290, 2022.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre- training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in neural information processing systems, vol. 33, pp. 12449–12460, 2020.
[9] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," arXiv preprint arXiv:1909.10681, 2019.
[10] J. Lee and W. Lee, "Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation," arXiv preprint arXiv:2108.11626, 2021.