Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 5, pp. 433-443, October 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.5.433

트랜스포머기반의 멀티모달 영상자막 생성요약

이민예 · 한성원†

고려대학교 산업경영공학과

Multi-Modal Abstractive Summarization based Transformer using Video Transcripts

Min Ye Lee · Sung Won Han

Department of Industrial Management Engineering, Korea University

In this paper, we propose a MASTF methodology, which is a Multimodal Abstractive Summarization based on Transformer. Neural network models applied in the field of generative summaries utilizing conventional multi-modals were techniques utilizing hierarchical attention based on circulating neural networks. Although transformers showed excellent performance in various natural language processing fields, including generative summaries, there were no cases of application in multimodal-based generative summaries. Thus, in this paper, we use transformers to improve the performance of multimodal image subtitle generation summary models. Transformer-based models outperform hierarchical attention-based models by 24.17% on ROUGE-L basis and 10.52% on combining speech and text.

Keywords: Transformer, Abstractive Summarization, Multi-Modal

1. 서 론

최근 영상 플랫폼은 양질의 대용량 콘텐츠를 확보하면서 단순한 영상 시청 플랫폼을 넘어 정보 검색 창구로 확대되고 있다. 따라서 방대한 영상 콘텐츠에서 사용자가 원하는 영상을 정확하게 찾고, 유사한 콘텐츠를 추천할 수 있는 검색 및 추천 기능의 필요성이 커져가고 있다(Song *et al.*, 2011; Wang *et al.*, 2012; Otani *et al.*, 2016; Torabi *et al.*, 2016). 영상 플랫폼의 검색 엔진은 영상의 제목, 설명(description), 그리고 태그의 텍스트를 분석한 후, 사용자가 입력한 검색어와 가장 적합한 영상을 추출한다. 하지만, 영상의 제목 및 설명은 제작자가 직접 입력하므로 주관적이며, 객관적인 검색이 어려워, 영상 플랫폼의 검색엔진 성능을 저하시킨다. 다른 영상이 동일한 태그를 갖는 경우에도 영상간의 미묘한 차이를 파악하기 어려우며 (Wang *et al.*, 2012), 텍스트가 미 기재된 영상일 경우 적절한 영상을 검색하는데 어려움이 있다. 이에 영상 자막(transcripts)을 요약하는 것은 영상에 대한 간략한 정보를 제공할 뿐만 아니라, 검색 키워드에 적합한 영상을 추출하는데 사용될 수 있다.

요약은 문서 생성 방식에 따라 추출요약(extractive summarization)과 생성요약(abstractive summarization)이 있다. 추출요약은 원문에 존재하는 문장을 그대로 가져오며, 생성요약은 문서의 내용을 압축하여, 새로운 문장을 만든다. 기존 생성요약 연구에는 순환신경망(recurrent neural networks: RNNs)기반의 시퀀스-투-시퀀스 방법, 트랜스포머 방법이 있다. 순환신경망 방법에 관한 선행연구는 Point-Generator Network(See *et al.*, 2017), Bottom-up attention(Gehrmann *et al.*, 2018) 등이 있으며, 트랜스포머 방법에 대한 선행연구는 MASS(Song *et al.*, 2019),

본 연구는 4단계 두뇌한국21에 의해 지원되었습니다. 이 논문은 2021년도 고려대학교의 지원을 받아 수행되었습니다. (No.K2107521, 제조 산업 데이터를 활용한 딥러닝 기반 이상 분할 탐지 기법 개발). 이 논문은 제1저자 이민예의 석사학위논문 내용을 포함하며 논문의 저작권은 대한산업공학회지에 있음.

† 연락저자 : 한성원 교수, 02841, 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3384, Fax : 02-929-5888, E-mail : swhan@korea.ac.kr

2021년 4월 19일 접수; 1차 2021년 6월 6일, 2차 2021년 7월 17일 수정본 접수; 2021년 8월 17일 게재 확정.

UniLM(Dong *et al.*, 2019), BART(Lewis *et al.*, 2019), PEGASUS (Zhang *et al.*, 2020), 그리고 ProphetNet(Yan, Y. *et al.*, 2020)를 참조한다. 일반적인 요약은 원문으로부터 핵심 내용이 담긴 짧고 간결한 글을 생성하는 단일모달리티 모델이지만, 영상 데이터는 자막, 음성, 그리고 연속적인 이미지 프레임으로 구성된 멀티모달(multi modal)데이터이다. Sanabria *et al.*(2018)은 멀티모달모형이 음성인식(automatic speech recognition), 기계번역(machine translation), 그리고 문서요약에서 단일모달리티 모형보다 뛰어난 성능을 가짐을 실험한다. 또한, 요리, 운동, 예술 그리고 실내외 활동에 대한 교육용 영상자료를 수집하여 음성 및 텍스트 데이터를 동반한 How2 멀티모달 데이터 셋을 공개하였다. 이러한 계기로 How2 데이터를 기반으로 영상에 수록된 자막, 음성, 그리고 이미지 정보를 융합하여 요약문을 생성하는 멀티모달기반의 영상 자막요약문을 생성하는 연구가 계속 이뤄지고 있다(Palaskar *et al.*, 2019; Khullar and Arora, 2020). Palaskar *et al.*(2019)은 상호보완적인 특성을 지닌 텍스트 모달리티와 영상 모달리티를 융합함으로써, 더 풍성하고 유창한 요약문을 생성할 수 있었다. 또한, Khullar and Arora (2020)는 음성 모달리티를 추가하였을 때, 유용한 정보를 줄 수 있다고 가정하고, 이를 실험한 결과 세 개의 모달리티를 융합하였을 시, 가장 우수한 성능을 보임을 실험하였다.

기존 멀티모달 자막생성요약에 대한 연구는 순환신경망기반의 계층적 어텐션(hieracical attention)기법을 적용한 연구가 주를 이룬다. Palaskar *et al.*(2019)은 텍스트 및 이미지 정보의 문맥 벡터를 계층적 어텐션을 통해 결합한 후, 디코더의 은닉상태로 입력하여 단일모달리티 모형을 뛰어넘는 성과를 기록하였다. Khullar and Arora(2020)은 텍스트 및 이미지 그리고 텍스트 및 음성을 결합한 두 문맥 벡터를 다시 어텐션하여 세 개의 모달리티의 계층적 어텐션(trimodal hieracical attention)을 제안한다. 트랜스포머(transformer)는 어텐션만으로 이루어진 인코더-디코더 모형이다(Vaswani *et al.*, 2017). 단어가 각 셀에 입력되는 순환신경망 구조를 탈피하고, 그라디언트 소멸(gradient vanishing)문제를 극복하여, 기계번역(Wang *et al.*, 2018), 질문-답변 생성(Rajpurkar *et al.*, 2016), 그리고 생성요약(Song *et al.*, 2019; Dong *et al.*, 2019; Raffel *et al.*, 2019; Lewis *et al.*, 2019)등 다양한 자연어 처리 분야에서 괄목할만한 성과를 보이고 있다. 트랜스포머의 어텐션은 키, 값에 입력되는 정보에 따라 원천-타겟 어텐션(Source-target attention)과 자기 자신을 참조하는 셀프 어텐션(Self attention)으로 구분된다. 원천-타겟 어텐션은 인코더-디코더 어텐션이라고도 불리며, 쿼리에 입력된 디코더의 은닉 상태에 대한 키, 값에 입력된 인코더의 은닉 상태(hidden state)를 참조한다. 멀티모달리티 분야에서는 원천-타겟 어텐션을 기반으로 키, 값에 참조하고자 하는 다른 모달 정보를 입력한다. 이를 크로스모달 어텐션(crossmodal attention)이라고 하며, 트랜스포머 기반의 구조에서 다양한 모달을 서로 융합하기 위한 방법으로 사용된다(Tsai *et al.*, 2019; Paraskevopoulos *et al.*, 2020). Tsai *et al.*(2019)는 트랜스포머 기반의 크로스모달 어텐션을 사용하여 비용이 많이 드는 멀티모달간의 정렬(alignment)을 수행하지 않고도 이미지, 오디오, 그리고 텍스트를 융합하여 감성인식(emotion analysis)분야에서 뛰어난 성능을 보였다. Paraskevopoulos *et al.*(2020)는 음성모달리티과 영상모달리티를 융합하여 음성인식(automatic speech recognition)분야에 적용한다.

트랜스포머는 다양한 자연어 처리 분야에서 순환 신경망보다 괄목한 성과를 보였음에도 불구하고, 멀티모달기반의 문서생성요약분야에 적용한 사례가 부족하다. 이에 본 연구에서는 트랜스포머의 크로스모달 어텐션을 활용하여 멀티모달 영상 자막 생성요약 방법론을 제안한다. 트랜스포머 기반의 모달 성능은 순환신경망의 성능보다 영상과 텍스트를 결합하였을 시, ROUGE-L기준 24.17%, 음성과 텍스트를 결합하였을 시, 10.52% 우수하였다. 영상 정보와 융합할 시 텍스트의 입력의 가중을 0.4만큼 주었을 때 성능이 가장 우수하였으며, 음성 정보와 융합할 시 텍스트의 입력의 가중이 0.8일 때 성능이 가장 우수하였다. 또한, 영상 정보와 융합한 모델의 성능이 음성 정보와 융합한 모델보다 더 우수하였다.

본 논문은 다음과 같이 구성되어 있다. 제2장에서는 문서생성요약, 멀티모달 생성모델에 대한 선행연구를 소개한다. 제3장에서는 연구에서 제안하는 방법론에 대하여 서술한다. 제4장에서는 How2데이터 셋을 가지고 제안한 모형을 실험한다. 마지막으로 제5장에서는 본 연구의 결론 및 활용방안을 서술한다.

2. 관련 연구

2.1 문서 생성 요약

기존 생성요약 연구에는 순환신경망(recurrent neural networks: RNNs)기반의 시퀀스-투-시퀀스 방법, 트랜스포머 방법이 있다. 순환신경망 방법에 관한 선행연구는 Point-Generator Network(See *et al.*, 2017), Bottom-up attention(Gehrmann *et al.*, 2018)등이 있으며, 트랜스포머 기반의 대한 선행연구는 MASS (Song *et al.*, 2019), UniLM(Dong *et al.*, 2019), BART(Lewis *et al.*, 2019), PEGASUS(Zhang *et al.*, 2020), 그리고 ProphetNet(Yan *et al.*, 2020)가 있다.

(1) 시퀀스-투-시퀀스

시퀀스-투-시퀀스(sequence-to-sequence, seq2seq)는 두 개의 순환신경망으로 구성된 아키텍처로 앞쪽은 인코더이고 뒤쪽은 디코더로 <Figure1>의 좌측 그림과 같다. seq2seq는 입력 시퀀스에 관한 새로운 시퀀스를 예측하는 언어 번역, 질문 생성, 그리고 텍스트 요약 과업에 사용된다. 고차원의 언어 데이터는 인코더에 의해 문맥 벡터(context vector)로 압축되며, 압축된 문맥 벡터는 디코더 RNNs의 초깃값으로 사용된다. 디코더

의 은닉상태는 각 시점에서 소프트맥스(softmax)활성화 함수를 가진 완전 연결층을 통과하여 어휘 사전에 등록된 단어들에 대한 확률분포를 출력한다. seq2seq는 순차적인 단어를 출력할 수 있는 모형으로 다양한 생성 요약 연구에 바탕이 되고 있다(Nallapati *et al.*, 2016; Paulus *et al.*, 2017; See *et al.*, 2017). 하지만, 입력 문장을 고정된 크기의 문맥 벡터로 압축하는 것은 정보의 손실을 유발시킨다. 그리고, 입력 문장이 길어질 경우, 그라디언트(gradient)가 소멸되어 요약 모델 성능이 저하되는 문제점이 있다(Werbos, 1990; Sutskever *et al.*, 2013).

(2) 어텐션 메커니즘

어텐션 메커니즘(attention mechanism)은 디코더의 매 시점마다 전체 입력 단어를 참조하며 <Figure1>의 우측 그림과 같다. 입력 단어 중 주의해야할 단어에 가중함으로써 seq2seq의 문제점을 해결한다. 예를 들어, 원천 문장 '킬리안 음바페는 아르헨티나와의 경기에서 후반 4분 만에 2골을 터뜨려 4대3으로 짜릿한 승리를 거두며 프랑스를 월드컵 8강에 진출시켰다.' 으로부터 '프랑스가 아르헨티나를 4대 3으로 꺾고 8강에 진출했다.' 라는 요약문을 생성할 때에, '꺾고' 라는 단어를 예측 시, '짜릿한 승리를 거두며' 단어에 집중을 해야 할 것이다.

어텐션 메커니즘은 인코더 내 주의해야 할 단어의 출력 상태는 디코더 내 예측하고자 하는 단어의 입력 상태와 유사하다고 가정한다. 이러한 유사성을 정렬점수(alignment score) e_{ij} 라고 하며, 정렬점수를 계산하는 방법은 Eq. (1)과 같다. Luong *et al.*(2015)은 정렬점수를 계산하는 법으로 내적(dot), 일반(general), 그리고 연결(concat)방식을 제시하였으며, 생성 요약 모델에서 일반과 연결방식이 주로 사용된다(Paulus *et al.*, 2017; See *et al.*, 2017). Eq. (2)와 같이 정렬함수를 통과한 정렬 벡터에 소프트맥스 함수가 적용되면 가중치 벡터가 생성된다. 가중치가 높을수록 단어 생성시 원문 내 주의를 주어야 할 단어를 찾을 수 있다. Eq. (3)은 은닉 상태 벡터와 가중치를 곱하여 더한 후 만들어진 문맥 벡터이다.

$$e_{ij} = \begin{cases} s_i^T h_j & \text{dot} \\ \mathrm{s_i^T W_{align} h_j} & \text{general} \\ (\mathrm{v_{align}})^{\mathrm{T}} \tanh(W_{align} s_i + U_{align} h_j) + b_{align} & \text{concat} \end{cases} \quad (1)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (2)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad (3)$$

Eq. (3)의 문맥 벡터 c_i와 직전 시점의 디코더의 은닉 상태 s_i 연결되어 완전 연결층을 통과한다. 이 과정은 Eq. (4)와 같다. 마지막으로, Eq. (5)에서 어휘 사전에 등록된 단어에 대한 확률분포를 출력한다. <Figure 2>는 디코더 예측 시점에 사용되는 문맥 벡터 c_i와 상기 벡터가 만들어지는 어텐션 메커니즘 과정을 나타낸다.

$$\tilde{s_i} = \tanh(W_s(c_i; s_i)) \quad (4)$$

$$P_{vocab,i} = softmax(W_{d2v}\tilde{s_i} + b_{d2v}) \quad (5)$$

See *et al.*(2017)은 출력 단어에 대한 확률 분포 $P_{vocab,i}$ 에다가 입력 단어에 대한 어텐션 확률 분포 α_i 를 더하는 포인터 생성 네트워크(pointer generator network)모형을 제안한다. 이는 단어 예측 시, 원문에서 중요한 단어들을 다시 재현하도록 하는

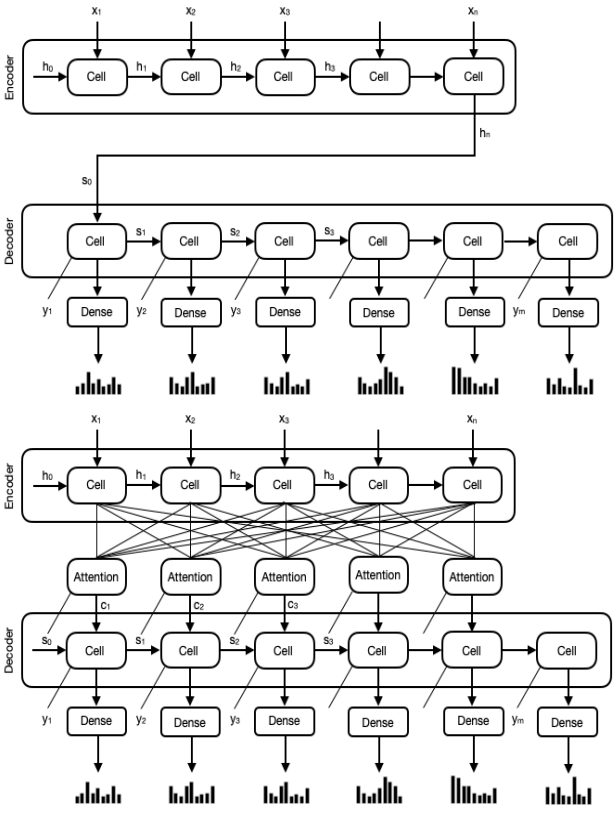

Figure 1. Seq2Seq Architecture and Attention Mechanism based on Seq2Seq

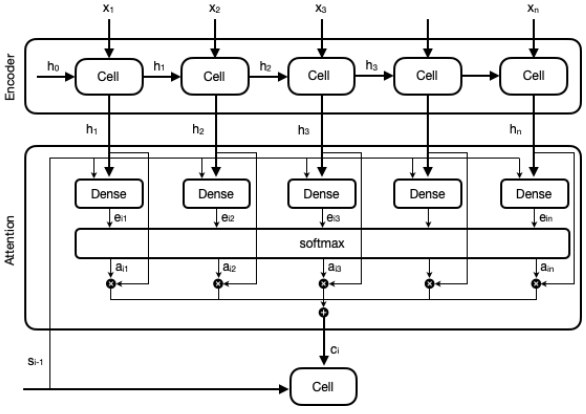

Figure 2. Attention Mechanism(Luong *et al.*, 2015)

역할을 한다. Gehrmann *et al.*(2018)은 입력 문장 중 요약문에 등장하는 단어에 대한 어텐션 확률 분포만 복사하는 컨텐츠 선택(content selector)모형을 제안한다. 이는 입력 문장 내 불필요한 부분까지 예측 하는 것을 방지한다.

(3) 트랜스포머

트랜스포머(transformer)는 어텐션 메커니즘만을 사용한 인코더-디코더 기반의 신경망이며, 구조는 <Figure 3>과 같다(Vaswani *et al.*, 2017). 길이가 L인 원천 문장은 임베딩 층을 통과한 후, d_{model}차원의 벡터가 된다. 트랜스포머는 문장의 시간적 구조를 추가하기 위하여 위치 임베딩(positional embedding)을 더하여 준다. 위치 임베딩은 Eq. (6)과 Eq. (7)과 같다.

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{6}$$

$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{7}$$

입력 임베딩은 쿼리입력x^Q, 키입력x^K, 그리고 값입력x^V으로 복사되며, 헤드(head)의 수만큼 쪼개지어 선형변환을 통과한 후 쿼리(Q), 키(K), 그리고 값(V)가 된다. 이는 Eq. (8)과 같다. 스케일드 점-곱 어텐션(scaled dot-product attention)은 Q와 K를 곱한 행렬을 스케일 조정 후 소프트맥스 함수를 통과하고, V와 곱하는 과정으로, Eq. (9)과 같다. 이 과정의 출력 벡터의 행의 합은 1로, 한 행은 특정 쿼리 단어가 입력 문장 내 다른 단어들과의 얼마나 유사한지 나타내는 확률이다. 인코더의 어텐션에서 Q, K, 그리고 V는 동일한 원천 문장이 입력되므로 셀프 어텐션(Self Attention)이라 하며, 이 과정의 출력 벡터는 어텐션 헤드(attention Head)이다.

$$\begin{cases} Q = x^Q W^Q \\ K = x^K W^K \\ V = x^V W^V \end{cases} \tag{8}$$

$$\mathrm{Attention(Q,K,V)} = \mathrm{softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{9}$$

멀티 헤드 어텐션(multi head attention)은 여러 개의 어텐션 헤드를 연결(concat)하여 선형 변환한 것이다. 스킵 연결을 통과한 원본 쿼리 입력과 원소별 덧셈을 한 후, 층 정규화(layer normalization)를 한다. 그 후 완전연결층을 통과한 출력을 스킵 연결과 층 정규화를 수행하여 쿼리 입력의 크기와 동일한 벡터를 출력한다. N개의 인코더의 층의 수만큼 반복되며, 마지막 인코더의 출력 벡터가 디코더의 입력으로 사용된다.

$$h^0 = embedding(x) \tag{10}$$

$$\widetilde{h}^l = LN(h^{l-1} + MHAtt(h^{l-1})) \tag{11}$$

$$h^l = LN(\widetilde{h}^l + FFN(\widetilde{h}^l)) \tag{12}$$

디코더의 마스크 멀티헤드 어텐션(masked multi-head attention)은 후속 타임스텝에 나오는 원소를 $-\infty$로 설정하며, 이는 미래 정보에 참조하지 않기 위함이다. 마스크 멀티헤드 어텐션의 출력 벡터는 멀티헤드 어텐션 과정의 쿼리(Query)로 입력된다. 키(Key), 값(Value)은 인코더의 출력이 사용되며, 출력 단어에 대한 입력 문장 내 주의해야 할 토큰을 알 수 있다. 인코더에서는 쿼리, 키, 그리고 값이 동일한 셀프 어텐션이였다면, 디코더에서는 키와 값이 원천문장이며, 쿼리가 타겟 문장인 원천-타겟 어텐션이다. 멀티헤드 어텐션 과정을 통과한 후, 선형 함수 및 소프트맥스 함수를 통과하여 전체 어휘 사전에서 예측하고자 하는 단어의 확률 분포를 추론할 수 있다.

트랜스포머는 어텐션만으로 인코더-디코더 구조를 구성하여, 단어가 각 셀에 입력되는 순환신경망 구조를 탈피하고, 그라디언트 소멸문제를 극복하였다. 이는 다양한 자연어 처리 분야에서 전이 학습 시 원천 도메인(source domain)을 학습하는 모형으로 사용되어 기계번역(Wang *et al.*, 2018), 질문-답변 생성(Rajpurkar *et al.*, 2016), 그리고 생성요약(Song *et al.*, 2019;

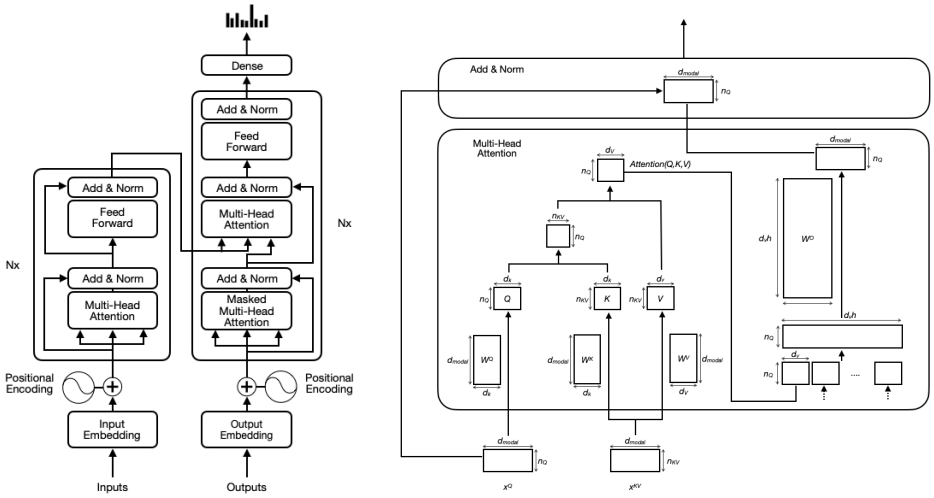

Figure 3. Transformer Architecture and Multi-Head Attention(Vaswani *et al.*, 2017)

Dong *et al.*, 2019; Raffel *et al.*, 2019; Lewis *et al.*, 2019)에서 괄목할만한 성과를 보이고 있다.

(4) 트랜스포머 기반의 생성 요약

버트(bidirectional encoder representation from transformers, BERT)는 트랜스포머의 인코더 구조를 기반으로 문장 내 랜덤한 토근을 마스킹하여 이를 예측하는 과업과 다음 문장을 예측하는 과업을 동시에 학습하는 양방향(bidirectional)모델이다(Devlin *et al.*, 2018). 단방향(unidirectional) 구조를 채택하여 좌측의 토큰만 참조하였던 ELMO와 GPT의 한계점을 극복한다. 하지만 버트의 양방향 모델은 자연어 이해(natural language understanding: NLU)과업에 적합하지만, 자연어 생성(natural language generation: NLG)에 적용하기 어렵다. 이에 트랜스포머 기반의 사전 학습 기법을 변형하는 다양한 생성요약 모형이 제안된다(Song *et al.*, 2019; Dong *et al.*, 2019; Lewis *et al.*, 2019; Zhang *et al.*, 2020; Yan *et al.*, 2020). MASS(MAsked Sequence to Sequence learning), UniLM(UNified pre-trained Language Model), BART(Bidirectional and Auto-Regressive Transformers), PEGASUS(Pre-training with Extracted Gap-sentences for Abstractive Summarization), 그리고 ProphetNet모형은 트랜스포머 기반의 생성요약 모형의 방법론이다.

MASS(Song *et al.*, 2019)은 원문 내 k개의 토큰을 마스킹 처리한 후, 이를 예측하는 전이학습 방법이다. 원천 문장의 토큰이 m개일 때, k는 $k \in \{1, m\}$의 범위를 가진다. k가 1일 경우, 버트의 학습 과정을 나타내며, k가 m일 경우, GPT의 학습을 나타낸다. 기가워드(gigaword)테스트 데이터를 사용하여 성능을 평가하였을 시, ROUGE-L 점수는 35.96이다. UniLM(Dong *et al.*, 2019)은 원문과 요약문을 연결한 후, 원문과 요약문 내 랜덤한 토큰을 마스킹 처리한다. 두 텍스트를 결합하여 인코더에 입력함으로써 원문과 요약문 사이의 관계를 학습하는 방법을 제안한다. MASS모델에 비해 ROUGE-L기준으로 7.08 정도의 성능 향상을 보였다. BART(Lewis *et al.*, 2019)는 토큰 마스킹(token masking), 토큰 삭제(token deletion), 텍스트 채우기(text infilling), 그리고 문장 순서 섞기(sentence permutation) 등을 통해 원천문장을 변형시킨 후, 디코더가 본래의 원문을 생성하는 모형이다. PEGASUS(Zhang *et al.*, 2020)은 하나의 문장 전체를 마스킹하여, 처리하는 갭 문장 생성(gap sentence generation)을 사용하는 사전 학습한다. PEGASUS는 BART보다 높은 성능을 보인다. ProphetNet(Yan *et al.*, 2020)는 디코더에서 하나의 타임스텝에 대한 예측이 아닌 n개의 단어를 예측하는 모형을 제안하며, 이는 가장 가까운 토큰과의 큰 상관성을 가지는 것을 방지한다.

2.2 멀티모달 생성모델

멀티모달은 음성인식(automatic speech recognition), 기계번역(machine translation), 그리고 문서요약 등 다양한 분야에 적용될 수 있으며, 단일 모달리티를 가진 모형보다 우수한 성능을 가진다(Sanabria *et al.*,2018).

(1) 멀티모달의 융합

멀티모달을 융합하는 방식에는 이른 융합(early fusion), 늦은 융합(late fusion), 그리고 중간 융합(intermidiate fusion)이 있다. 딥러닝(deep neural network)은 고차원의 이질적(heterogeneous)데이터를 저차원의 벡터로 변환시킴으로써, 결합 표현(joint representation layer) 또는 공유 표현층(shared representation layer)이라 하는 동일한 잠재 공간 안에 표현(representation)하는 중간 융합 방법으로 분류된다. 이 결합 표현층을 학습하는 방법으로 계층적 어텐션(hierarchical attention)과 교차모달 어텐션(cross modal attention)방법이 있다.

(2) 계층적 어텐션

계층적 어텐션은 서로 다른 두 모달의 인코더(encoder)로부터 출력된 벡터를 어텐션하여 멀티모달의 결합 표현(joint representation)을 학습하는 기법이다(Libovicky and Helcl, 2017). <Figure 4>는 텍스트, 영상, 그리고 음성에 대한 계층적 어텐션을 나타낸다. Eq. (13)는 k번째 모달리티의 어텐션 층에서 출력된 문맥 벡터와 디코더 타임 스텝의 직전 은닉 상태를 기반으로 계산된 어텐션 스코어이다. 소프트 맥스 함수를 통과하여 가중치가 되며 이는 Eq. (14)과 같다. 텍스트와 음성의 결합은 Eq. (15)와 같으며, 이와 마찬가지로 텍스트와 영상의 결합도 구할 수 있다. 이 두 결합을 다시 어텐션 하여 세 모델에 대한 어텐션 메커니즘을 구축한다(Khullar and Arora, 2020).

$$e_i^{(k)} = (v_{align})^T \tanh\left(W_{align} s_i + U_{align}^{(k)} c_i^{(k)} \right) \tag{13}$$

$$\beta_i^k = \frac{\exp(e_i^{(k)})}{\sum_{n=1}^{N} \exp(e_i^{(n)})} \tag{14}$$

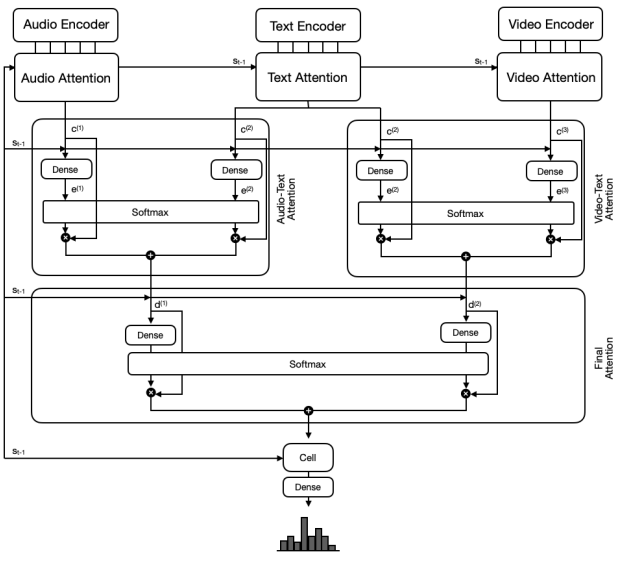

Figure 4. Hierarchical Attention

$$d_i = \sum_{k \in \{\text{audio,text}\}}^{N} \beta_i^{(k)} U_c^{(k)} c_i^{(k)} \quad (15)$$

(3) 교차모달 어텐션

교차모달 어텐션은 스케일드 점-곱 어텐션(scaled dot-product attention)을 사용하여 타겟 모달리티를 쿼리에 입력하고, 다른 모달리티를 키, 값에 입력한다. 타겟 모달리티와 상관성이 높은 다른 모달리티 내 요소를 찾음으로써 두 모달리티를 결합할 수 있다(Tsai *et al.*, 2019; Paraskevopoulos *et al.*, 2020). Tsai *et al.*(2019)은 멀티모달 트랜스포머(Multimodal Transformer: MulT)를 제안하여 여러 모달의 정렬(alignment)이 없이 감성 인식 분야에서 뛰어난 성능을 보였다. MulT의 핵심 모듈은 교차 어텐션(cross-modal attention)이다. Paraskevopoulos *et al.*(2020)은 음성 인식을 위해 음성 모달과 영상 모달을 사용하여 이를 교차 어텐션한다.

3. 방법론

본 논문에서는 트랜스포머 기반의 멀티모달 생성요약 모형(Multimodal Abstractive Summarization based Transformer: MASTF)을 제안한다. MASTF모형은 임베딩(embedding) 및 특징 추출, 인코더(encoder), 멀티모달의 융합(fusion), 그리고 디코더(decoder)단계로 구성된다. 전체적인 MASTF연구 방법은 <Figure 5>와 같다. 임베딩은 입력 자막을 벡터로 표현하며, 특징 추출 부분에서는 영상 또는 음성 모달리티의 특징을 추출한다. 인코더(Text Encoder)는 입력 시퀀스로부터 문맥 벡터를 추출하며, 셀프 어텐션 층과 피드 포워드 신경망으로 구성된다. 멀티모달의 융합에서는 교차모달 어텐션을 기반으로 텍스트 모달리티와 영상 및 음성 모달리티가 융합된다. 텍스트 인코더의 출력벡터와 멀티 모달리티를 융합한 교차모달 어텐션의 출력벡터가 결합하여 디코더에 입력된다. 이는 <Figure 5>의 융합(Fusion) 블록에 있으며, 융합 블록의 출력은 인코더의 최종 출력값으로 디코더의 멀티 헤드 어텐션(Multi-Head Attention)블록에서 키와 값으로 사용된다. 디코더의 현재 상태인 쿼리를 기반으로 키와 값으로 입력된 인코더 최종 벡터에서 중요한 정보를 획득한 후, 가장 적합한 다음 단어를 예측한다. 완전 연결층을 통과한 벡터가 실제 단어를 생성하기 위하여 디코더의 최종 단계에서는 선형 층 및 소프트맥스 함수를 두어 단어 사전 내 모든 단어에 대한 확률 분포를 출력한다. 가장 높은 확률을 가진 단어가 다음 단어로 예측이 되며, 이 과정은 순차적으로 이루어진다.

3.1 임베딩 및 특징 추출

텍스트 모달의 임베딩(embedding)층은 각 단어를 임베딩 크기 길이의 벡터로 변환하는 룩업 테이블(Lookup table)이다. 임베딩 층에서 학습되는 가중치의 수는 임베딩 크기 길이와 어휘 사전 크기의 곱이다. 입력 층은 원문 글자 수×배치사이즈 크기의 정수 시퀀스 텐서를 임베딩 층으로 전달하며, 임베딩 층은 원문 글자 수×배치사이즈×임베딩 차원 수 크기의 텐서를 출력한다. 트랜스포머는 순차적 구조를 가지지 않으므로 위치 인코딩(positional encoding)을 임베딩과 합한다.

영상의 특징을 추출하는 과정은 <Figure 6>과 같다. 영상은 연속적인 이미지 프레임으로 변환되며, 매 16번째 이미지 프레임을 추출한다. ImageNet데이터를 사전 학습한 ResNet-50 (He *et al.*, 2016)을 사용하여 각 이미지 프레임에서 2048크기의 특징을 추출한다.

음성 인식 및 신호처리를 위한 오픈소스 Kaldi(Povey *et al.*, 2011)를 사용한다. 16kHz 원본 신호에서 10ms 프레임 시프트한 후, 25ms의 윈도우 크기를 사용하여 40차원의 filter bank특징을 추출하고, 3차원의 peatch특징을 연결한다. 또한 비디오 당 화자의 변동성을 추가하기 위해 Cepstral 평균 및 분산 정규화(cepstral mean and variance normalization: CMVN)을 적용한다. 음성은 샘플링 수×배치사이즈×특징 차원 수(43)크기의 텐서를 가진다.

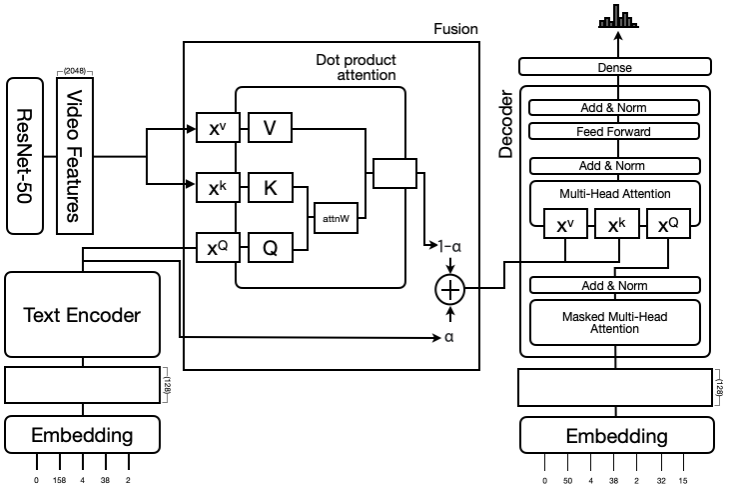

Figure 5. Overview of MASTF Methodology

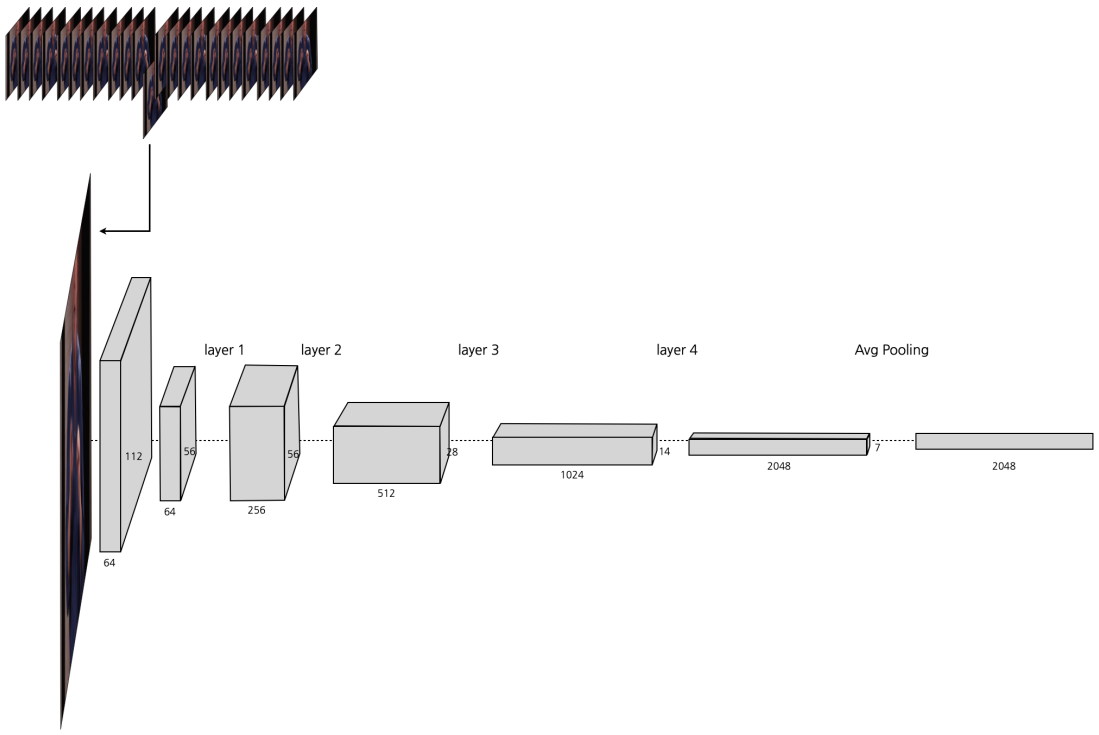

Figure 6. Image Feature Extraction using ResNet-50

3.2 인코더

임베딩 및 특징 추출이 되면 각 모달에서 시퀀스 길이×배치 사이즈×특징 차원 수를 가진 행렬이 출력된다. 텍스트의 특징 차원 수는 128이며, 영상은 2048, 그리고 음성은 43차원이다. 텍스트 임베딩은 순차 구조를 추가하기 위하여 위치 임베딩과 합하여진 후, 트랜스포머의 인코더에 입력된다. 상기 행렬은 각각 쿼리입력, 키입력, 그리고 값입력으로 복사되어 인코더 내 멀티헤드 어텐션(multi head attention)과정에 입력된다. 쿼리입력, 키입력, 그리고 값입력은 선형함수를 통과하여 쿼리(Q), 키(K), 그리고 값(V)이 된다. 인코더의 멀티헤드 어텐션은 Q, K, 그리고 V가 모두 동일한 셀프 어텐션(Self Attention)메커니즘이다. 원본 쿼리 입력과 원소별 덧셈을 한 후, 층 정규화(layer normalization)를 한다. 그 후 완전연결층을 통과한 동일하게 스킵 연결 및 층 정규화를 수행한다. 인코더의 출력은 쿼리 입력과 동일한 벡터를 가진다.

3.3 멀티모달의 융합 및 디코더

텍스트 인코더로부터 출력된 행렬은 입력쿼리에 저장되며, 영상 또는 오디오로부터 출력된 행렬은 키-값입력에 사용된다. 영상 모달리티를 β라 하고, 텍스트 모달리티를 α라 한다면, 영상 모달리티는 입력쿼리에 사용되어 쿼리 $Q_\alpha = X_\alpha W_{Q_\alpha}$가 되며, 텍스트 모달리티는 키-값입력에 사용되어 키$K_\beta = X_\beta W_{K_\beta}$와 값 $V_\beta = X_\beta W_{V_\beta}$이 된다. 가중치가 $W_{Q_\alpha} \in R^{d_\alpha \times d_k}$, $W_{K_\beta} \in R^{d_\beta \times d_k}$, $W_{V_\beta} \in R^{d_\beta \times d_k}$ 이며, 모달리티의 입력은 $X_\alpha \in R^{T_\alpha \times d_\alpha}$, $X_\beta \in R^{T_\beta \times d_\beta}$ 일 때, 교차모달 어텐션 $Y_\alpha := CM_{\beta-\alpha}(X_\alpha, X_\beta) \in R^{T_\alpha \times d_v}$은 Eq. (16)과 같다. 교차모달 어텐션 과정을 거치면서, 텍스트의 각 입력 원소는 상관성이 높은 영상 혹은 음성 모달리티의 요소를 찾고 결합한다.

$$Y_\alpha = softmax(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}}) X_\beta W_{V_\beta} \qquad (16)$$

디코더는 요약문의 후속 타임스텝에 나오는 단어에 주의를 주지 않기 위해 해당 원소를 $-\infty$로 설정한다. 이 과정을 마스크 멀티 헤드 어텐션(masked multi-head attention)이라고 하며, 출력 벡터는 이후의 멀티 헤드 어텐션 과정의 입력 쿼리로 사용된다. 교차모달 어텐션의 출력은 멀티헤드 어텐션 과정에 키-값 입력에 사용되여 출력 단어를 예측하기 위해 입력 모달리티에 대한 중요도를 알 수 있다. 텍스트 인코더로부터 출력된 행렬에 α계수를 곱하며, 교차모달 어텐션의 출력에는 $1-\alpha$계수를 곱한다. 마지막으로 선형 과정을 거치고, 소프트맥스 함수를 통과하여 전체 어휘 사전에 대한 확률 분포를 추론할 수 있다.

4. 실험 결과

4.1 데이터 전처리

데이터는 Sanabria *et al.*(2018)이 공개한 How2 데이터셋을

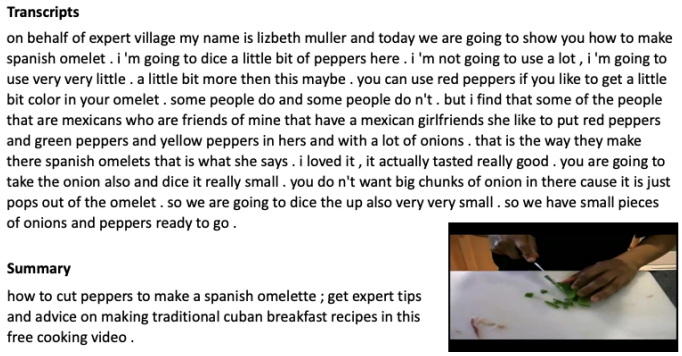

Figure 7. How2 Dataset

사용한다. <Figure 7>은 데이터 셋 중의 일부로 영상, 음성, 영상 자막(transcripts), 그리고 요약문(summary)으로 구성된 How2 데이터셋을 나타낸다. 영상 자막(transcripts)은 화자의 음성을 문자 데이터로 변환한 것으로 상세한 내용을 담고 있는 반면, 요약문(summary)는 전체 영상에 대한 추상화된 개요를 담고 있다. 예를 들어, 요약문에 언급된 'cut' 및 'Cuban breakfast recipe' 등은 영상 자막에 존재하지 않는다. 텍스트 모달리티와 영상 모달리티 정보를 상호보완함으로써, 더 풍성하고 유창한 요약문을 생성할 수 있다.

영상을 이미지로 변환하였을 시, 최대 길이는 691개이며, ResNet-50을 사용하여 각 이미지 프레임에서 2048크기의 특징을 추출한다. 오디오의 최대 샘플링 수는 31,892 이다. 원문의 최대 단어의 수는 1,212이며 요약문 내 최대 단어 수는 87개이다. 학습, 검증, 그리고 테스트 데이터의 수는 각각 12798, 520, 그리고 127개이며, 영상의 시간은 총 300시간이다.

4.2 하이퍼파라미터 선택

<Table 1>는 실험으로부터 최종 선택된 신경망의 초 매개변수(hyperparameter)이다. 초 매개변수는 학습 과정이 시작되기 전에 사용자가 설정해주는 값을 의미하며, 은닉 노드의 수, 학습 진도율(learning rate), 반복횟수(epoch), 최적화 방법, 그리고 배치사이즈(batch size)가 있다. 학습진도율은 0.0004로 지정하였으며, 반복횟수는 50으로 설정하며, 검증 데이터가 20회 이상 향상되지 않을 경우 학습이 조기 종료된다. 최적화 방법은 아담(Adaptive Moment estimation : Adam)방법을 사용한다. 배치사이즈는 64이다.

Table 1. Hyperparameter Configuration

	Transformer	RNN
Architecture	Encoder - Transformer Fusion - Crossmodal Attention Decoder - Transformer	Encoder - LSTM Fusion - Hierachical Attention Decoder - LSTM
Hidden Size	128	128
Learning-rate	0.0004	0.0004
Training epochs	50	50
Patience	20	20
Optimizer	Adam	Adam
Batch Size	64	4
No. Layers	4	
No. Heads	4	

4.3 평가 지표 및 모델 비교

멀티모달 요약모델의 성능을 평가하기 위해 루지(Recall-Oriented Understudy for Gisting Evaluation :ROUGE)와 블루(Bilingual Evaluation Understudy :BLEU)점수를 사용한다. ROUGE-N은 모델 요약문의 N그램이 정답 요약문과 겹치는 비율을 의미하며, Eq. (17)과 같다.

$$ROUGE-N=\frac{\sum_{S\in Reference}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in Reference}\sum_{gram_n\in S} Count(gram_n)} \quad (17)$$

ROUGE-1은 유니그램(Unigram), ROUGE-2는 바이그램(Bigram), ROUGE-N은 N그램, 그리고 ROUGE-L은 모델 요약문과 정답 요약문 사이에 가장 긴 공통 서브 시퀀스가 차지하는 비율을 나타낸다. BLEU는 모델 요약문이 정답 요약문과 겹치는 비율을 의미하며, Eq. (19)과 같다. Eq. (18)의 BP(brevity panalty)는 모델 요약문 c가 정답 요약문의 길이 r보다 작을 경우 패널티를 부여한다.

$$BP=\begin{cases}1 & \text{if } c>r\\ e^{(1-r/c)} & \text{if } c\le r\end{cases} \quad (18)$$

$$BLEU=BP\cdot\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (19)$$

<Table 2>는 텍스트에 대한 영상 또는 음성 데이터를 참조하였을 시, 모델의 성능을 보여준다. α는 텍스트 인코더로부터 출력된 행렬의 계수이며, $1-\alpha$는 크로스모달 어텐션으로부터 출력된 행렬의 계수이다. Driven은 모델에서 유도된 α 값을 사용한 경

Table 2. ROUGE-N, BELU on Test Data of Models based on Transformer

		ROUGE			BLEU
		1	2	L	
Transformer(Video-Text)	$\alpha=0.8$	41.7645	21.0421	39.9461	10.0609
	$\alpha=0.6$	42.3152	22.1436	41.1485	17.8997
	$\alpha=0.4$	42.1757	22.3684	41.1866	18.588
	$\alpha=0.2$	40.9141	20.8752	40.6521	16.4246
	Driven	**43.3903**	**23.109**	**42.5416**	**19.0915**
Transformer(Audio-Text)	$\alpha=0.8$	**41.5403**	20.5006	**41.1485**	**17.8997**
	$\alpha=0.6$	41.4444	20.7665	40.1941	15.5216
	$\alpha=0.4$	41.0353	**21.1242**	40.6723	17.5197
	$\alpha=0.2$	39.0213	18.2166	37.8464	12.3199
	Driven	39.0368	19.6406	37.5474	15.7011

Table 3. ROUGE-N, BELU on Test Data of Models based on RNN and Transformer

	ROUGE			BLEU
	1	2	L	
First 80 tokens(Text Only)	22.3428	5.3609	21.8737	3.1
RNN(Text Only)	15.8475	2.5922	18.4049	0.399
RNN(Video-Text)	33.6737	17.6605	34.2591	**29.9351**
RNN(Audio-Text)	37.5794	18.98	37.2308	23.9267
RNN(Video-Audio-Text)	38.4739	18.4949	39.3017	18.3485
Transformer(Text-Only)	41.5003	20.5349	39.7585	17.2905
Transformer(Video-Text)	**43.3903** (+28.85%)	**23.109** (+30.85%)	**42.5416** (+24.17%)	19.0915 (-36.22%)
Transformer(Audio-Text)	41.5403 (+10.54%)	20.5006 (+8.01%)	41.1485 (+10.52%)	17.8997 (-25.18%)

우를 의미한다. α가 1에 가까울수록 텍스트 정보 입력 비중이 높아진다. 영상 정보와 융합할 시 α가 0.4일 때 성능이 가장 우수하였으며, 음성 정보와 융합할 시 α가 0.8일 때 성능이 가장 우수하였다. α계수를 학습하였을 시, 영상 정보와 융합한 모델의 성능이 음성 정보와 융합한 모델보다 더 우수하였다.

<Table 3>는 순환 신경망과 트랜스포머 기반의 멀티모달 생성요약 모델의 성능을 비교한다. 트랜스포머는 순환신경망보다 병렬 처리, 긴 입력 문장에서의 정보 손실 문제, 그리고 그라디언트 소멸 문제 등을 해결함으로써 트랜스포머 기반의 모달 성능은 순환신경망의 성능보다 영상과 텍스트를 결합하였을 시, ROUGE-L기준 24.17%, 음성과 텍스트를 결합하였을 시, 10.52% 우수하였다.

<Table 4>는 순환 신경망과 트랜스포머 기반의 모델에서 추출된 요약본과 정답 요약본을 비교한다. 정답 요약본에 따르면 본문의 내용은 스크럽 크림을 이용한 피부관리법에 대한 영상이다. 텍스트 기반의 순환 신경망 모형은 영상의 핵심 키워드인 스크럽 크림이라는 단어가 있으나 스크럽 크림을 만드는 방법이라 요약한다. 음성, 또는 영상과 융합된 순환 신경망 모형은 피부 관리에 대해 언급을 하나, 본 내용과 관련이 없는 단어가 언급이 된다. 이미지와 융합한 트랜스포머 기반의 모형은 스크럽을 사용한 피부관리 법이라는 정답 요약문의 내용과 가장 비슷하게 요약을 한다.

we're going to be removing the scrub and you most always want to use a nice warm towel on your client to remove the scrub, and that just to get to really remove all the granules that are left over and a nice warm towel usually helps in doing so. and a granular scrub like this one is really going to help sloughed off surface layers of the skin but also, you know, if someone is experiencing any flakiness of the skin, if they've recently had a facial treatment with a booster, like a glycolic peel or any other peel, sometimes you experience a little bit of flakiness a few days after. using a gentle scrub like this one will help to remove those flaky areas. again, you want to be extremely, extremely gentle when using this product . but usually, you just get a softer and more resurface texture to your skin. and again, it's also clarifying which is brightening. and again, using this every two to three times a week is really going to help boost your skin care regimen and your skin care goals and needs .

Table 4. Summary Generated by Different Models for Document of G33aqW3DLpc

Reference	wash off a facial scrub cream gently and note any flaky skin areas . learn tips for removing home facial scrubs from an esthetician in this free spa treatment video
RNN(Text Only)	learn how to make a facial scrubs from an esthetician in this free spa treatment video
RNN(Video-Text)	the skin with tips from a manicurist in this free massage therapy video
RNN(Audio-Text)	's facial products in this free dental health video from our medical care
RNN(Video-Audio-Text)	get rid of facial skin care with tips from an aesthetician in this free pet care video get rid of nail polish facial safely by using a facial scrub with information from an experienced aesthetician
Transformer(Text-Only)	clean treatment with acid free products for cleaning and apply acid to clean your skin learn about spa treatment for dust framed with tips from a licensed esthetician in this free skin care video
Transformer(Video-Text)	facial skin is useful for applying oily skin learn how to apply a facial scrubs for a facial skin treatment with expertise from a licensed medical aesthetician in this free skin care video
Transformer(Audio-Text)	clean hair with dry cleaners by using a towel to clean the skin before cleaning products inside the skin learn more about acrylic nails with tips from an aesthetician in this free skin care video

Table 5. Summary Generated by Different Models for Document of FzQPg4aqNYc

Reference	how to serve apple rice salad ; get professional tips and advice from an expert chef on making some unique, fresh salads in this free cooking video .
RNN(Text Only)	our expert cook how to make a apple rice salad in this free recipe dessert video taught by our expert chef loves you
RNN(Video-Text)	how to use this and more in this free video on home organization and maintenance .
RNN(Audio-Text)	learn about the history of food, with tips from a professional chef in this free video about how to make chicken fajitas
RNN(Video-Audio-Text)	our expert cook how to make a apple rice salad in this free recipe dessert video taught by our expert chef loves you
Transformer(Text-Only)	how to add sugar to your gourmet coffee cake recipe get expert tips on baking homemade dessert recipes in this free cooking video
Transformer(Video-Text)	how to brown off rice for apple rice salad get professional tips and advice from an expert chef on making some unique fresh salads in this free cooking video
Transformer(Audio-Text)	learn from an expert about how to mix the ingredients for making chocolate chip oatmeal cookies in this free cooking video on making clouser minnows for making jewelry

<Table 5>의 정답 요약본에 따르면 사과 쌀 샐러드를 만드는 법에 대한 영상이다. 이미지 기반의 트랜스포머 모형은 사과 쌀 샐러드라는 핵심 단어가 존재하여 비교적 적합한 요약문을 보여준다. 또한, 순환신경망 기반 모형 중 이미지 및 음성 모두 융합한 모형에서도 사과 쌀 샐러드를 만드는 법이라고 요약한다.

hi ! this is rich buccola here in new york city and today on behalf of expert village, i'm going to show you how to make an apple crisp rice salad. we're back. i just took our salad out of the refrigerator. we let it go about another half an hour. everything fused really nice. let's show you how we plate it. i put some arugula. if you have any knife of lettuce or leafy greens is fine. i have my mixing cup. what i do is just put a whole bunch in here and give it a good pat like this. it actually takes the shape of the cup. i'm going to actually just spoon it right on to our salad like so . what i 'll do then is just push some freshly ground black pepper on this all around . i usually take an arugula leaf and just put it right on top . stick it in . that 's how i serve that .

5. 결론 및 활용방안

본 논문에서는 트랜스포머 기반의 멀티모달 영상자막 생성요약 방법론을 제안한다. 기존 멀티모달을 활용한 생성요약 분야에 적용된 신경망 모형은 순환 신경망 기반의 계층적 어텐션을 활용한 기법이었다. 트랜스포머는 생성요약을 포함한 다양한 자연어 처리 분야에 우수한 성능을 보임에도 멀티모달 기반의 생성요약 분야에 적용한 사례가 없었다. 이에 본 논문에서는 트랜스포머를 사용하여 멀티모달 영상자막 생성요약 모형의 성능을 향상시킨다. 계층적 어텐션을 사용한 모델보다 트랜스포머 기반의 모델이 ROUGE-L기준 24.17%, 음성과 텍스트를 결합하였을 시, 10.52% 우수하였다.

MASTF모형은 트랜스포머 기반의 멀티모달 생성요약 모형이다. 텍스트, 영상, 그리고 음성 등 멀티모달 데이터를 모델링하여 요약문을 생성한다. 영상에 대한 핵심적인 정보를 제공할 뿐만 아니라 검색 시스템에서 영상의 반환을 위한 판단 기준으로 사용될 수 있다. 본 연구에서는 텍스트와 영상, 그리고

텍스트와 음성을 융합한 실험을 하였으나, 세 가지 모달에 대한 융합은 비교하지 않는다. 향후 연구에서는 세 가지 모달리티를 융합하는 방식을 실험하고, 다양한 데이터에서 트랜스포머가 성능에 미치는 영향을 탐구하는 것이다.

참고문헌

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H. (2019), *Unified Language Model Pre-training for Natural Language Understanding and Generation*.

Gehrmann, S., Deng, Y., and Rush, A. M.(2018), *Bottom-Up Abstractive Summarization*, EMNLP.

He, K., Zhang, X., Ren, S., and Sun, J. (2016), Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 770-778.

Khullar, A. and Arora, U. (2020), *MAST: Multimodal Abstractive Summarization with Trimodal Hierarchical Attention*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020), *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation*, Translation, and Comprehension.

Libovický, J. and Helcl, J. (2017), *Attention Strategies for Multi-Source Sequence-to-Sequence Learning*.

Luong, T., Pham, H., and Manning, C. D. (2015), *Effective Approaches to Attention-based Neural Machine Translation*.

Nallapati, R., Zhou, B., Santos, C. D., Gülçehre, Ç., and Xiang, B. (2016), *Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond*, CoNLL.

Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., and Yokoya, N. (2016), *Learning Joint Representations of Videos and Sentences with Web Image Search*.

Palaskar, S., Libovický, J., Gella, S., and Metze, F. (2019), *Multimodal Abstractive Summarization for How2 Videos*, ACL.

Paraskevopoulos, G., Parthasarathy, S., Khare, A., and Sundaram, S. (2020), *Multiresolution and Multimodal Speech Recognition with Transformers*.

Paulus, R., Xiong, C., and Socher, R. (2018), *A Deep Reinforced Model for Abstractive Summarization*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Vesel, K. (2011), The Kaldi speech recognition toolkit, *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Raffel, C., Shazeer, N. M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020), Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *J. Mach. Learn. Res.*, **21**(140), 1-67.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016), *SQuAD: 100, 000+ Questions for Machine Comprehension of Text*, EMNLP.

Sanabria, R., Çağlayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F.(2018), *How2: A Large-scale Dataset for Multimodal Language Understanding*.

See, A., Liu, P. J., and Manning, C. D. (2017), *Get To The Point: Summarization with Pointer-Generator Networks*, ACL.

Song, J., Yang, Y., Huang, Z., Shen, H. T., and Hong, R. (2011), Multiple Feature Hashing for Real-time Large Scale Near-duplicate Video Retrieval, In *Proceedings of the 19th ACM international conference on Multimedia*.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. (2019), *MASS: Masked Sequence to Sequence Pre-training for Language Generation*, ICML.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013), On the Importance of Initialization and Momentum in deep Learning, In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, PMLR, **28**(3),1139-1147.

Torabi, A., Tandon, N., and Sigal, L. (2016), *Learning Language-Visual Embedding for Movie Understanding with Natural-Language*.

Tsai, Y. H., Bai, S., Liang, P., Kolter, J. Z., Morency, L., and Salakhutdinov, R. (2019), Multimodal Transformer for Unaligned Multimodal Language Sequences, *Proceedings of the conference. Association for Computational Linguistics*, Meeting, 6558-6569 .

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017), Attention is All you Need.

Wang, M., Hong, R., Li, G., Zha, Z., Yan, S., and Chua, T. (2012), Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification, *IEEE Transactions on Multimedia*, **14**(4), 975-985.

Wang, Y., Wu, J., and Hoashi, K. (2019), Multi-Attention Fusion Network for Video-based Emotion Recognition, *International Conference on Multimodal Interaction*.

Werbos, P. J. (1990), Backpropagation Through Time: What it does and How to do it, In *Proceedings of the IEEE*, **78**(10), 1550-1560.

Yan, Y., Qi, W., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. (2020), *ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training*.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020), *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*, ICML.

저자소개

이민예 : 한동대학교 컴퓨터공학과에서 2016년 학사학위를 취득하고, 고려대학교 산업경영공학부 석사과정에 재학 중이다. 연구 분야는 딥러닝, 수요 예측, 그리고 자연어 처리이다.

한성원 : 고려대학교 산업 시스템 정보 공학과에서 2003년 학사학위를 취득하였다. Georgia Institute of Technology에서 2006년 Operation Research 석사학위, 2007년 Statistics 석사학위, 2010년 Mathematics 석사학위를 취득하고 2010년 Industrial Engineering과 Statistics 박사학위를 취득하였다. University of Pennsylvania, Department of Biostatistics and Epidemiology에서 Post-doctoral Researcher(2010.07~2012.06), Hoffmann-La Roche Inc., Department of Non-clinical Safety에서 Post-doctoral Fellow(2012.07~2013.08), New York University, Department of Population Health에서 Research Scientist (2013.08~2015.12), New York University, Department of Population Health에서 Senior Research Scientist (2016.01~2016.02)을 역임하고, 2016년부터 고려대학교 산업경영공학부 교수로 재직하고 있다. 연구분야 중 방법론 분야는 probabilistic graphical model, network analysis, deep learning 등이 있으며, 응용 분야로는 바이오 의료, 소재 정보학, 품질 모니터링, 텍스트 마이닝이 있다.