이학석사학위논문

복잡한 비디오 질의응답 문제를 위한 계층 구조의 멀티모달 인코더와 사전훈련

Hierarchical Multi-modal Encoder and Pre-training for Long-term Dependencies and Reasoning in the Video Question Answering task

2021 년 8 월

서울대학교 대학원

협동과정 뇌과학전공

임 형 석

복잡한 비디오 질의응답 문제를 위한 계층 구조의 멀티모달 인코더와 사전훈련

Hierarchical Multi-modal Encoder and Pre-training
for Long-term Dependencies and Reasoning
in the Video Question Answering task

지도교수 장 병 탁

이 논문을 이학석사 학위논문으로 제출함

2021 년 8 월

서울대학교 대학원

협동과정 뇌과학전공

임 형 석

임형석의 이학석사 학위논문을 인준함

2021 년 8 월

위 원 장	이 인 아
부위원장	장 병 탁
위 원	곽 노 준

초록

비디오 질의응답(Video Question Answering) 문제는 비디오의 시각 정보와 언어 정보를 이용하여 실환경에서의 다양한 관계를 효과적으로 이해할 수 있는 능력을 요구하는 과제로, 인간 수준의 통합된 지능을 가져야 할 인공지능 에이전트를 평가하기에 효과적인 과제이다. 하지만 지난 몇 년간의 비디오 질의응답 관련 연구는 짧은 길이의 비디오 클립에서 질의 응답으로 진행되었다. 짧은 길이의 비디오 뿐만 아니라 길고 복잡한 비디오의 질의응답을 위해서는 장기의존성 (long-term dependency)과 높은 수준의 추론 능력을 사용하여 해결해야한다. 따라서 본 연구에서는 복잡하고 긴 비디오 클립 내의 장기 의존성 문제를 해결하면서 어려운 질의응답 문제를 위한 추론 능력을 향상시키기 위해 계층적 구조를 가진 새로운 멀티모달(Multi-modal) transformer 구조와 대조 학습(Contrastive learning)을 이용한 새로운 pre-training task를 제안한다. 세 단계의 계층 구조로 이루어진 multi-modal transformer를 제안해 장기 의존성을 인코딩하고 비디오 장면에 대한 이해력을 향상시키며, 질문과 관련이 높은 비디오 표현을 사용하여 비디오의 전반적인 맥락을 학습하고 정답을 추론한다. 또한 Video-Subtitle Matching (VSM) task와 새롭게 제안한 Video-QA Matching (VQAM) task를 이용해 비디오의 표현 뿐만 아니라 복잡한 질문에 대해 정답을 고를 수 있도록 효과적으로 pre-training을 한다. 그 중 제안한 VQAM은 질의와의 어텐션 기법과 대조 학습을 활용하여 질의응답을 위한 추론 능력을 향상시킨다. 제안된 전체 프레임워크의 평가를 위해 네 가지 어려움 단계를 가진 한국 드라마 질의응답 데이터셋인 DramaQA 데이터셋을 이용하여 실험하여 그 결과 어려운 수준의 질의응답에서 좋은 결과를 보인 것을 확인하고, 본 모델을 통해 효과적인 비디오 표현을 학습할 수 있다는 것을 기존 모델들과의 비교와 ablation

study를 통해 확인하였다.

주요어: 딥러닝, 비디오 질의응답, 멀티모달 학습, 사전훈련

학번: 2018-22635

목차

표 목차

그림 목차

제 1 장 서론

인간은 실환경에서 다른 존재와 소통하거나 상호 작용하기 위하여 시각적인 정보와 함께 언어, 소리, 촉각, 제스쳐 등의 다양한 정보를 활용한다. 이처럼 오늘날의 컴퓨터 비전(Computer vision) 분야에서는 단순히 시각적인 정보만을 다루는 문제에서 벗어나 음성이나 문자 등 다른 형태의 정보도 같이 활용하여 지금까지 풀기 힘들었던 문제들을 다루고 있다. 그중에서도 비디오에서 장면 이해에 대한 연구는 실환경과 비슷한 공간에서 대화나 사건을 통해 이야기가 전개되거나 사람 또는 물체 간의 시공간적인 관계를 파악할 수 있는 시각적, 언어적 정보를 모두 이해해야 하기 때문에 모델의 능력을 인간의 인지 과정을 모사할 수 있도록 발전시키는 데에 큰 도움을 주고 있다. 특히 시각 정보만을 이용한 물체 인식(Object detection) (Yilmaz et al., 2006)과 행동 인식(Action recognition) (Ji et al., 2012)과 같은 연구 뿐만 아니라 비디오의 시각 정보 및 언어 정보를 모두 활용할 수 있는 text-to-video retrieval (J. Xu et al., 2016), video captioning (Krishna et al., 2017)과 같은 다양한 연구로 발전하고 있어서 반드시 필요한 연구 중 하나이다.

그 중 비디오 질의응답(Video Question Answering, VideoQA) 문제는 비디오 클립의 내용과 관련된 질의를 보고 여러 개의 선택지 중 질문에 알맞는 답을 고르는 과제로, 비디오 안의 시각 정보와 언어 정보를 통합하여 이해하면서 복잡한 추론도 수행하여야 하기 때문에 현재 어려움을 겪고 있다. 이 문제를 해결하기 위해 지금까지의 많은 연구들은 (D. Xu et al., 2017; Jang et al., 2017; Gao et al., 2018; Choi et al., 2020) LSTM 기반의 인코더(Encoder)-디코더(Decoder) 구조를 사용하는 접근법을 채택해왔다. 비디오 프레임에서 convolutional neural networks (CNNs)를 이용해 visual feature를 추출하고 자막 입력을 워드 임베딩으로 변환한 후에 각 feature를

LSTM 인코더를 거친 후 합쳐주는 late fusion 방법과 두 feature를 합친 후에 LSTM 인코더를 거치는 early fusion 방식이 대표적이다. 그러나 이런 LSTM 계열의 순환 신경망은 짧은 길이의 비디오 클립에서 효과적이고 좋은 성능을 보이지만, 비디오의 길이가 길어지고 장면의 변환이 잦아지는 경우에는 장기 의존성(long-term dependency)을 크게 고려하지 못하기 때문에 긴 길이의 비디오 클립에서는 적절하지 않다. 지금까지의 비디오 질의응답 문제를 위한 데이터셋들 (K.-M. Kim et al., 2017; Jang et al., 2017; Lei et al., 2018)은 1분 미만의 짧은 길이를 가진 비디오 클립으로 구성되었거나 질의응답에 요구되는 실제 비디오의 길이가 짧았기 때문에 한 비디오 클립 내의 장면 간의 장기적인 의존성을 고려하지 않아도 괜찮았지만, 최근에 발표된 데이터셋 (Choi et al., 2020)에서는 1 – 10분 길이의 비디오 클립들과 어려운 질문을 포함하고 있기 때문에 장기적인 의존성과 더불어 복잡한 추론 능력을 요구한다. 또한 질의 자체가 복잡하거나 여러 단계의 추론을 거쳐야 하는 경우 비디오 질의응답 문제는 더 어려워진다. 이를 해결하기 위한 시도로 외부 메모리를 사용한 연구 (Zeng et al., 2017; Gao et al., 2018)가 있다. 외부 메모리에는 비디오 입력을 LSTM으로 인코딩한 feature가 저장되어 있고, LSTM으로 인코딩된 질의와의 어텐션 기법을 이용해 관련 정보를 검색하고 메모리를 업데이트 해준다. 이과 같은 메모리 업데이트를 반복하여 비디오 입력으로부터 질문과 관련된 여러 번의 추론을 하는 효과를 낼 수 있다. 하지만 LSTM 기반의 인코더를 사용하기 때문에 질의의 복잡한 의미를 모두 담아내지 못하거나 긴 길이의 비디오 클립의 전체 맥락(global context)에서 질문을 이해하지 못한다는 한계가 존재한다.

본 논문에서는 이와 같은 비디오 클립 내의 장기 의존성 문제를 해결하면서 어려운 질의응답 문제를 위한 추론 능력을 향상시키기 위해 계층적 구조를 가진 새로운 멀티모달(Multi-modal) transformer 구조와 대조 학습(Contrastive learning)을 이용한 새로운 pre-training task를 제안한다. 먼저 일반적으로 많이 사용하는 LSTM 계열의 인코더 대신 세 단계의 계층 구조로 이루어진 multi-modal transformer를 제

안한다. 첫 번째 단계의 cross-modal transformer에서는 비디오 프레임과 해당 자막을 입력으로 받아 통합시켜 local context를 가지는 멀티모달 embedding을 만드는 역할을 한다. 두 번째 단계인 temporal transformer에서는 한 비디오 클립의 모든 비디오 프레임에 대해 local 멀티모달 embedding들을 하나로 합쳐서 장기 의존성을 가지는 global context feature를 얻는다. 그리고 마지막 단계에선 다시 한번 cross-modal transformer를 적용하여 global context상에서 서로 정보를 주고 받아 전체 비디오 클립과 자막간의 의미적인 정렬을 할 수 있도록 보장한다. 본 프레임워크를 pre-training 하기 위한 방법으로 Video-Subtitle Matching (VSM)과 새롭게 제안한 Video-QA Matching (VQAM)을 사용하여 pre-training task를 진행한다. VSM에서는 자막이 어떤 비디오 클립에 해당하는 지에 대한 global alignment와 비디오 클립 내에서 어느 시간 대에 위치하고 있는 지에 대한 local alignment에 대해 학습하면서 비디오와 자막 간의 시간적인 정렬을 맞춰주는 역할을 하고 학습과정에서 질의응답에 필요한 비디오의 부분이 어디인지 추출해내는데 역할을 한다. 제안한 VQAM에서는 먼저 multi-modal transformer로 얻은 통합된 비디오 표현과 질의와의 어텐션 기법을 적용해 질의에 해당하는 답을 추론하는 멀티모달 표현을 얻고, 인코딩된 질의응답 쌍에 대해 정답인 경우 답을 추론한 멀티모달 표현과의 유사도가 크도록, 반대로 오답인 경우에 멀티모달 표현과의 유사도가 작도록 대조 학습을 진행하여 질의응답의 문제를 위한 추론 능력을 향상시킨다. 본 프레임워크를 통해 나온 representation의 성능을 평가하기 위해 네 가지 어려움 단계를 가진 한국 드라마 질의응답 데이터셋인 DramaQA를 사용하여 평가를 진행하며, 기존 모델들과의 비교를 통해 그 효과를 증명한다.

제 2 장 관련 연구

2.1 시각 질의응답(Visual Question Answering)

언어를 기반으로 한 시각 정보를 이해하는 것은 인간 수준의 통합된 지능을 가져야 할 인공지능 에이전트에게 중요한 능력이다. 시각 질의응답 문제(Visual Question Answering, VQA) (Antol et al., 2015)는 그림 2.1과 같이 이미지나 영상이 있을 때 해당 상황에 맞는 질의가 존재하고 그 질의에 적절한 답을 생성하거나 몇 가지의 답변 선택지에서 정답을 고르는 작업으로, 실환경의 시각과 언어를 모두 다뤄야하는 인공지능 에이전트의 능력을 평가하기 위한 과제로 적합하다. 이런 시각 질의응답 문제를 성공적으로 수행하기 위해서는 물체 인식(Object detection) (Redmon et al., 2016), 행동 인식(Action recognition) (Ji et al., 2012)과 같은 높은 수준의 장면 이해 뿐만 아니라 지식 기반 추론(Knowledge-based reasoning) (Socher et al., 2013), 상식 추론(Commonsense reasoning) (Davis & Marcus, 2015) 등과 같이 장면에서의 복잡한 추론 능력이 요구된다.

시각 질의응답 연구는 데이터의 종류에 따라 크게 두 가지 형태로 나누어서 진행된다. 하나는 질의에 대한 정답을 예측하기 위해 공간적인 정보가 집중된 단일 이미지에 질문이 포함된 형태의 데이터를 이용하여 질의응답을 수행하는 것 (Tapaswi et al., 2016)이고, 다른 하나는 시간적인 정보에 자막과 같은 언어 정보가 추가된 비디오와 같은 형태의 데이터에서 질의 응답을 수행하는 것 (Lei et al., 2018)이다. 후자의 연구를 그림 2.1과 같은 비디오 데이터를 이용한 질의응답(Video Question Answering, VideoQA) 과제라고 한다. 비디오 데이터를 이용한 질의응답은 데이터로부터 시각 정보 뿐만 아니라 언어 정보도 같이 제공되기 때문에 이를 추가적으로 다루기 위해 시각과 언어 정보가 통합된 멀티모달 표현을 학습하는 과정과 학습

된 멀티모달 표현을 이용하여 질의에 맞는 정답을 추론하는 과정이 필요하다. 특히 이미지와는 달리 시공간적 정보를 담고 있는 대규모 데이터를 다루기 때문에 크고 복잡한 모델을 사용하게 된다.

2.2 멀티모달 학습(Multi-modal Learning)

인간의 경험은 여러 감각이 함께 작용하는 복합적인(Multi-modal) 형태로 이루어진다. 멀티모달 데이터란 다양한 형태(Modality)의 정보로 이루어져 통계적 특성이 구분되는 데이터를 뜻한다. 예를 들어 이미지 시퀀스(Sequence), 텍스트 자막(Subtitle), 그리고 음성 데이터들로 함께 구성되어 있는 비디오 데이터가 있다. 이처럼 다양한 형태의 데이터 특징을 효과적으로 학습하기 위한 방법을 멀티모달 학습(Multi-modal learning)이라고 한다. 다양한 멀티모달 데이터 중 하나인 비디오 데이터는 현재 가장 활발하게 진행되고 있는 인공지능 연구 과제 (Ngiam et al., 2011) 중 하나이다.

인공지능에서 멀티모달 학습으로 변수 차원이 각기 다른 다양한 형태의 데이터셋을 활용하는 것은 데이터에 대한 정보량을 증가시키고 모델의 성능을 높이는 역할을 한다. 그러나 서로 다른 모달리티를 결합하는 것은 실제로 다양한 잡음과 모달리티간의 충돌로 인해 어려움을 겪고 있다. 그래서 최근에는 멀티모달 학습을 위해 다양한 형태의 데이터를 통합(fusion)하는 방법이 중요한 주제로 자리 잡게 되었다.

초창기의 멀티모달 퓨전(multimodal fusion)은 early fusion LSTM (Snoek et al., 2005)과 late fusion LSTM (Lazaridou et al., 2015)으로 LSTM의 입력이나 출력 단계에서 각 모달리티 벡터(Vector)를 더하거나 concatenate하여 통합을 수행하였고, 이러한 방식은 단일 모달리티를 사용했을 때보다 더 좋은 성능을 보였다. 최근에는 attention을 이용한 멀티모달 퓨전 (Hori et al., 2017; Zadeh et al., 2018) 방법이나 NLP 분야에서 많이 사용되고 있는 Transformer 구조를 이용한 멀티모달 퓨전 (Tsai

et al., 2019; Li et al., 2020) 방법이 제안되었다. Transformer 구조를 이용한 멀티모달 퓨전은 별도의 alignment 없이 멀티모달 데이터를 활용하여 기존의 LSTM을 기반으로 한 방식의 한계였던 데이터의 장기적인(Long-term) 의존성을 고려하지 못했던 한계를 극복할 수 있다는 장점이 있지만, 대용량 데이터와 엄청난 연산량을 필요로 하는 점이 단점이 있다.

2.3 자기지도 학습(Self-supervised Learning)

자기지도 학습(Self-supervised Learning)은 비지도 학습(Unsupervised learning)의 일종으로 레이블(Label)이 없는 데이터를 기반으로 한 학습이며 자기 스스로 학습 데이터에 대한 학습을 수행하는 방식을 말한다. 사전훈련(Pre-training)은 라벨링 되지 않은(Unlabeld) 대용량 데이터로 모델을 미리 학습시키는 것으로, multi-layered perceptron(MLP)에서 weight와 bias를 잘 초기화 시키는 방법으로 제안되었다. 이를 통해서 여러 개의 hidden layer도 효율적으로 훈련시킬 수 있다. 자기지도 학습에서는 사용자가 pretext task라는 pre-training을 위한 새로운 문제를 정의하며 정답도 사용자가 직접 정해주게 된다. 컴퓨터 비전에서는 많은 수의 pretext task들이 제안되었는데, 예를 들어 이미지에 왜곡(Distortion) (Dosovitskiy et al., 2015)이나 회전(Rotation) (Gidaris et al., 2018)을 주는 것과 같은 pseudo labeling 방법 (Vincent et al., 2010)이나 autoencoder를 사용하여 손상된 이미지를 복원 (Zhang et al., 2016) 시키거나 Colorization하는 방법 (Pathak et al., 2016; Zhang et al., 2017) 들이 있다.

인공지능 모델로 하여금 pretext task를 학습하게 하여 데이터에 대해 높은 수준으로 이해를 높일 수 있게하고, 모델을 pre-training한 후에 최종적으로 달성하고자 하는 문제인 downstream task로 전이 학습(Transfer learning)을 하는 것이 자기 지도 학습의 핵심 개념이다. 자기 지도 학습은 BERT (Devlin et al., 2018) 같은 language modeling이나 generative model에서 자주 사용되었는데, 현재는 컴퓨터 비전(Computer vision)이나 음성 인식(Speech recognition), 로보틱스(Robotics) 등 다양한 분

야에서도 사용되고 있다. 하지만 최근 자기 지도 학습 연구는 대조 학습 기반의 학습 모델 (He et al., 2020; Chen et al., 2020)을 이용하여 기존 방식들보다 뛰어난 성능을 얻고 지도 학습 모델과의 격차를 줄이고 있다.

2.4 대조 학습(Contrastive Learning)

대조 학습(Contrastive learning) (Carreira-Perpinan & Hinton, 2005)은 데이터의 특징 표현을 효율적으로 추출해내는 학습 방법으로 최근 다양한 분야의 머신 러닝(Machine learning) 및 딥러닝(Deep learning) 연구에서 사용되고 있다. 대조 표현 학습의 주요 목표는 현재의 데이터와 매칭이 되는 데이터의 특징 벡터를 가깝도록, 현재의 데이터와 다른 데이터에 대해서는 특징 벡터가 멀어지도록 학습하는 것을 의미한다. 예를 들어 그림 2.2처럼 anchor와 같은 이미지라고 생각되는 이미지를 positive example, 다른 이미지라고 생각되는 이미지를 negative example이라고 하자. Convolutional neural networks(CNNs) (Krizhevsky et al., 2012)와 같은 매핑 함수로 이미지의 feature를 추출한 후 metric function을 정의해서 feature들간의 유사도를 측정하여 positive example pair의 유사도는 크도록, negative example pair는 유사도는 작도록 학습시키는 방법이 대조 학습이다. Contrastive loss는 식 (2.1)과 같이 softmax의 log loss 식으로 표현할 수 있다.

$$\mathcal{L}_q = -\log \frac{\exp\left(q \cdot k_+/\tau\right)}{\sum_{i=0}^{K} \exp\left(q \cdot k_i/\tau\right)} \tag{2.1}$$

Video clip

Scripts

No, No, not you. Pretty Oh Haeyoung.
Since there were two Oh Haeyoungs in one class, she was the Pretty Oh Haeyoung
I was the …
Hey, Regular Oh Haeyoung. Isn`t it your turn? Erase the board!

QA

Q. Why is Haeyoung1 holding a whiteboard eraser?

1. Haeyoung1 is holding a whiteboard eraser to pass to Haeyoung2.
2. Haeyoung1 is holding a whiteboard eraser to throw it.
3. Haeyoung1 is holding a whiteboard eraser to play with it.
4. Haeyoung1 is holding a whiteboard eraser to buy it.
5. Haeyoung1 is holding a whiteboard eraser to clean whiteboard.

그림 2.1: Video QA 예시

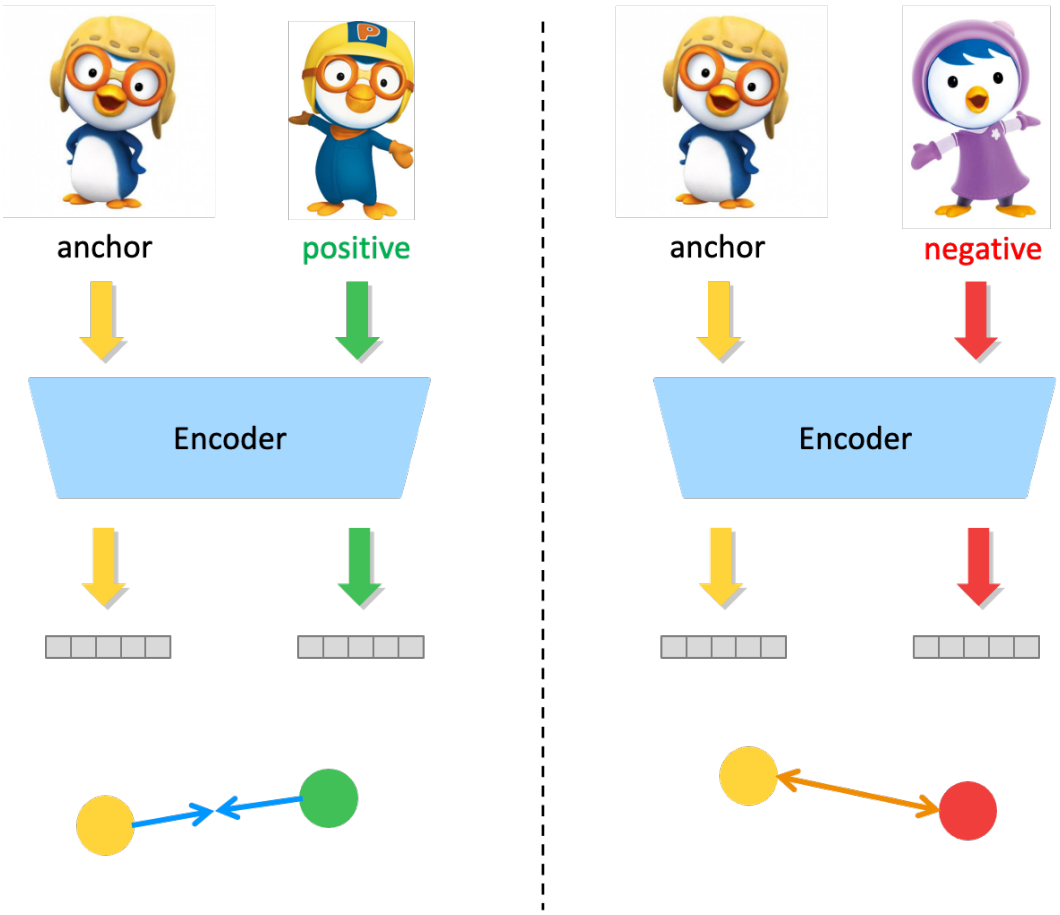

그림 2.2: Contrastive learning 학습 예시

제 3 장 방법론

본 장에서는 먼저 비디오 질의응답(Video Question Answering, VideoQA) 과제를 해결하기 위한 전반적인 과정을 구체적으로 정의하고, 이를 효과적으로 해결하기 위해 본 논문에서 제안한 프레임워크의 각 모듈과 pre-training task, 그리고 fine-tuning 학습 과정과 implementation details에 대해 자세히 설명한다.

3.1 문제 정의

비디오 질의응답 데이터셋에서 (Lei et al., 2018; Choi et al., 2020) 입력은 다음과 같다: (1) 질의 Q; (2) 질의에 해당하는 비디오 클립 $V = \{v_t\}_{t=1}^{N_v}$ (N_v는 비디오 클립의 프레임 수); (3) 비디오의 타임스탬프에 해당하는 자막 $S = \{s_t\}_{t=1}^{N_s}$ (N_s는 비디오 자막의 문장 수); (4) 5개의 정답 후보 $\{a_i\}_{i=1}^{5} \in \mathcal{A}$ ($\mathcal{A}$는 정답 공간).

이 네 가지의 데이터를 입력으로 받아 모델 $\mathcal{P}_\theta$을 학습하고, 그 결과물로 질문에 대한 정답 $\hat{a}$을 찾는 것이 본 비디오 질의응답 문제의 목표이다. 이는 다음의 식 (3.1)과 같이 표현할 수 있다.

$$\hat{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \mathcal{P}_\theta(a \mid Q, V, S) \tag{3.1}$$

3.2 모델 구조

본 논문에서 제안하는 모델은 비디오의 입력들을 인코딩해주는 embedding module, 다른 두 비디오 입력을 하나의 통합된 표현으로 나타내주게 하는 multi-modal transformer, 그리고 통합된 멀티모달 표현과 질문을 이용하여 정답을 유추할 수 있는 표현을 얻게 하는 어텐션 기반의 reasoning module의 총 세가지 모듈로 이루어져 있다. 모델의 전반적인 구조는 그림 3.1과 같다.

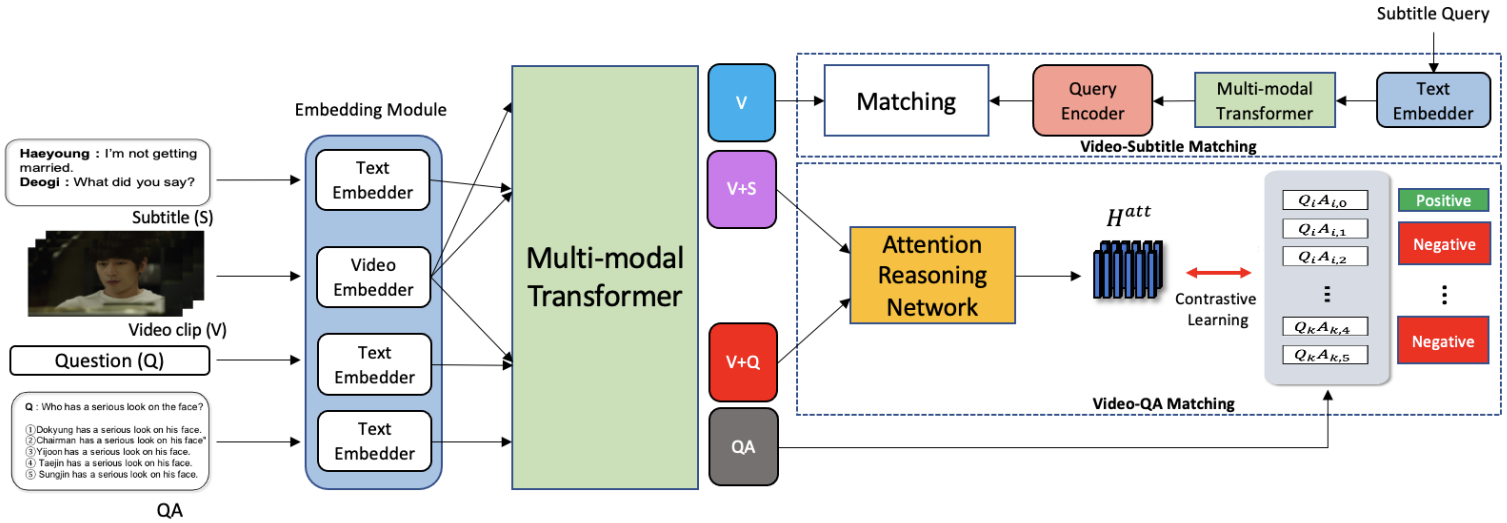

그림 3.1: 전체 모델의 pre-training 구조

3.2.1 Embedding Module

비디오 자막과 질의응답 같은 텍스트 표현은 pre-trained BERT (Devlin et al., 2018)모델을 사용하여 embedding 한다. 먼저 한 질의에 포함된 자막의 문장들을 모아 WordPieces (Wu et al., 2016)를 이용해 tokenize한 후 각 token에 대한 임베딩을 얻는다. 그리고 난 후 positional embedding과 함께 layer normalization을 거쳐서 text embedding $T_{s_t} = \{t_{s_t}^k\}_{k=1}^{N_t} \in \mathbb{R}^{N_t \times h}$ (N_t는 자막 문장들에서 token 개수, h는 hidden 사이즈)을 얻는다.

비디오 표현은 ImageNet 데이터셋 (Deng et al., 2009)으로 pre-trained 된 ResNet-101 (He et al., 2016)과 Kinetics 데이터셋 (Kay et al., 2017)으로 pre-trained 된 I3D (Carreira & Zisserman, 2017)를 사용하여 얻도록 구성한다. 질의에 해당하는 비디오 클립을 입력으로 ResNet과 I3D에서 얻은 feature들을 concatenate 한 후 fully connected layer를 통하여 텍스트 임베딩과 같은 차원으로 투영해 layer normalization을 해준다. 즉, video embedder를 통해 얻은 visual embedding은 $V_{st}^{emb} \in \mathbb{R}^{N_O \times h}$과 같이 표현될 수 있다. ($N_O$는 질문에 해당하는 프레임 개수)

Embedding module을 거치면 결과적으로 자막(S)과 질의(Q), 응답(A)에 대한 text embedding과 video frame embedding(V)을 얻는다.

3.2.2 Multi-modal Transformer

Multi-modal transformer에서는 서로 다른 형태의 데이터를 하나의 멀티모달 표현으로 나타내는 것 뿐만 아니라 hierarchical transformer 구조로 비디오와 같은 시계열 데이터의 문제점인 long-term dependency modeling을 해결하는 역할을 한다. Multi-modal transformer는 두 종류의 transformer를 이용해 세 단계의 계층 구조로 구성하였고, 전반적인 구성은 그림 3.2과 같다.

먼저 첫 번째 단계인 cross-modal transformer (Tan & Bansal, 2019)는 비디오 클립에서의 비디오 프레임들과 자막들을 local context 관점에서 정렬시켜 텍스트

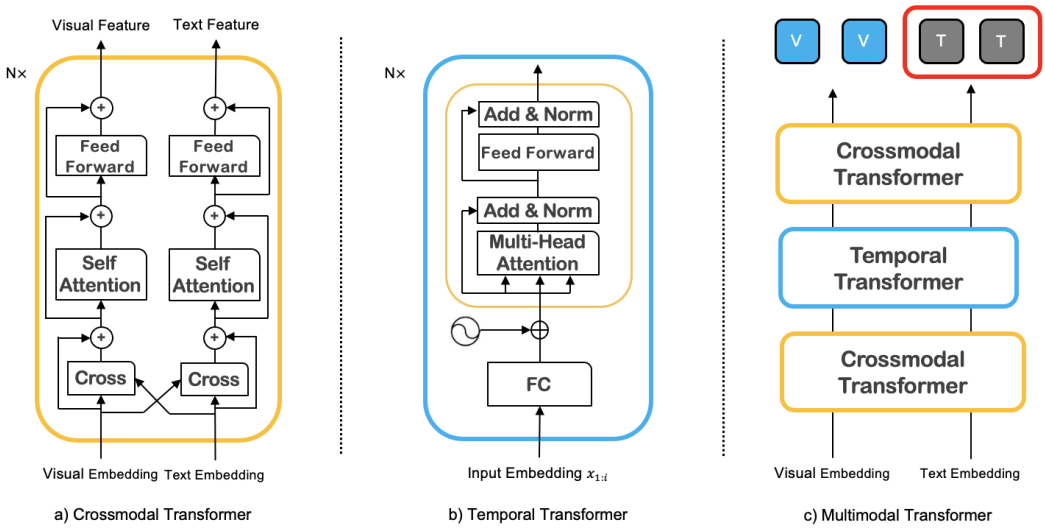

그림 3.2: Multi-modal Transformer

token과 비디오 프레임을 합성하는 인코더이다. Cross-modal transformer는 각 입력에 대해 bi-directional cross-modal attention layer, self-attention layer, 그리고 feed forward layer로 구성된 transformer를 N_L번 stack한 multi-layer transformer 구조로 이루어져있다. 그 결과로 video frame과 subtitle token에 대한 feature를 얻을 수 있다.

$$V_{s_t}^{cross}, W_{s_t}^{cross} = f_{cross}(V_{s_t}^{emb}, T_{s_t}) \tag{3.2}$$

두 번째 단계에서는 temporal transformer를 거친다. Temporal transformer는 비디오 클립의 모든 시간대에서 cross-modal feature를 합성함으로써 비디오의 전체적인 맥락(global context)을 이해할 수 있는 표현을 얻는 것이 목표이다. Temporal transformer는 (Vaswani et al., 2017)에서 제안된 기본 transformer 모델의 encoder 구조를 가지고 있다. Temporal transformer의 결과물로 contexutalized 된 video embedding과 text embedding 를 얻을 수 있다.

$$V^{temp} = f_{temp}(V^{cross}, 0) \tag{3.3}$$

$$W^{temp} = f_{temp}(0, W^{cross}) \tag{3.4}$$

마지막으로 두 번째 단계에서 얻은 video embedding과 text embedding V^{temp}, W^{temp}에 cross-modal transformer를 다시 적용하여 global context에서의 정보 교환과 정렬을 맞춰준다. Cross-modal transformer의 결과로 나온 두 feature 중에 text feature $\tilde{W} \in \mathbb{R}^{N_t \times h}$를 최종 multi-modal representation으로 사용한다.

3.2.3 Attention-based Reasoning Network

어텐션 기반의 reasoning module에서는 질의와 비디오간의 밀접한 관계를 내용을 인코딩하고, 질의의 정답에 가까운 형태로 representation을 mapping 시켜주는 것이 목적이다. 그래서 질의와 관련된 multi-modal representation 부분에 더 집중시키기 위하여 multi-modal transformer에서 얻은 video-subtitle 표현 벡터 $\tilde{W} \in \mathbb{R}^{N_{t_s} \times h}$와 video-question 표현 벡터 $\tilde{Q} \in \mathbb{R}^{N_{t_q} \times h}$ 간의 관계에 attention을 적용하여 정답을 추론하는 멀티모달 표현을 얻는다. 이를 위해 먼저 두 표현 벡터의 dot-product와 softmax 함수를 통해 attention distribution을 계산한다.

$$e^t = score(\tilde{W}, \tilde{Q}) = W^t Q \tag{3.5}$$

$$\alpha^t = softmax(e^t) \in \mathbb{R}^{N_{t_s} \times N_{t_q}} \tag{3.6}$$

그 다음 attention value를 계산해서 최종적인 attention 표현 H^{att}을 얻는다.

$$W^{att} = \alpha_t \tilde{Q} \tag{3.7}$$

$$H^{att} = W^{att} W_Q \in \mathbb{R}^{N_{t_s} \times h} \tag{3.8}$$

3.3 사전 훈련(Pre-training) Task

본 논문에서는 앞에서 제안한 모델을 효과적으로 학습하기 위해서 두 가지 pre-training task를 적용한다. 각 모듈에 대한 설명은 다음과 같다.

3.3.1 Video-Subtitle Matching Task

Video-Subtitle Matching (VSM)은 같은 시간대에 해당하는 비디오와 자막을 매칭시킴으로써 멀티모달 간의 시간적인 정렬을 장려하여 multi-modal encoder의 성능을 향상시키고, 질의응답을 위해 질의와 관련된 장면을 찾아내는데 효과적인 역할을 할 수 있는 학습 방법이다. VSM은 (Li et al., 2020)의 방식을 따라 pre-training 한다.

본 task에서는 비디오 입력의 local alignment와 global alignment를 학습한다. Local alignment에서는 비디오 내에서 자막이 해당되는 부분의 시작과 끝 부분의 프레임을 예측하도록 학습하고, global alignment에서는 비디오에서 랜덤으로 샘플한 자막이 올바른 비디오 클립에 매칭되도록 학습한다. 입력은 비디오 클립에서 랜덤으로 샘플된 자막 문장 s_q, 전체 비디오 클립 v, 그리고 샘플 된 자막 문장을 제외한 나머지 비디오 자막 문장 s_q이다. 비디오 클립과 샘플된 자막을 각각 embedding module과 multi-modal transformer를 거쳐 얻은 결과는 다음 식 (3.10)와 같다.

$$V^{all} = f_{trans}(V^{emb}, 0) \in \mathbb{R}^{N_v \times h} \tag{3.9}$$

$$W_{s_q} = f_{trans}(0, W_{s_q}^{emb}) \tag{3.10}$$

subtitle feature W_{s_q}에는 추가적으로 (Lei et al., 2020)에서 사용한 self-attention layer, linear layer와 layer normalization로 이루어진 query encoder를 사용하여 최종적으로 query vector $q \in \mathbb{R}^h$를 얻는다.

Local alignment에서는 query vector q와 video feature V^{all}의 dot product로 각 frame마다의 local matching score를 구한다. Matching score에 두 개의 학습 가능한

1D convolution filter와 softmax layer를 적용하여 시작 프레임과 끝 프레임에 대한 확률 벡터 $p_{start}, p_{end} \in \mathbb{R}^{N_v}$를 구할 수 있다. 그리고 cross-entropy loss를 이용하여 다음의 식 (3.11)과 같이 학습하여 자막에 해당하는 비디오 클립의 시작과 끝 부분의 프레임을 예측한다.

$$\mathcal{L}_{\text{local}} = -\mathbb{E}(\log\left(\mathbf{p}_{start}\left[y_{start}\right]\right) + \log\left(\mathbf{p}_{\text{ed}}\left[y_{ed}\right]\right)) \tag{3.11}$$

Global alignment에서는 query vector q와 video feature V^{all}의 코사인 유사도(Cosine similarity)를 max-pooling하여 다음의 식과 같이 global matching score S_{global}를 구한다:

$$S_{\text{global}}\left(s_q, \mathbf{v}\right) = \max\left(\frac{\mathbf{V}^{all}}{\|\mathbf{V}^{all}\|}\frac{\mathbf{q}}{\|\mathbf{q}\|}\right) \tag{3.12}$$

그 후 hinge loss를 적용한 학습을 위하여 positive와 negative video-query pair를 구성한다. Positive pair (s_q,v)에 대해서 v나 s_q를 같은 mini-batch의 다른 샘플로 대체하여 두 개의 negative pair (s_q,$\hat{v}$), ($\hat{s_q}$,v)를 만들어준다. 즉, global alignment의 training loss와 VSM의 최종 loss는 식 (3.13)와 같다.

$$\begin{aligned}
&\mathcal{L}_h\left(S_{\text{pos}}, S_{neg}\right) = \max\left(0, \delta + S_{neg} - S_{\text{pos}}\right) \\
&\mathcal{L}_{\text{global}} = -\mathbb{E}_D\left[\mathcal{L}_h\left(S_{\text{global}}\left(s_q, \mathbf{v}\right), S_{\text{global}}\left(\hat{s}_q, \mathbf{v}\right)\right)\right. \\
&\left.+\mathcal{L}_h\left(S_{\text{global}}\left(s_q, \mathbf{v}\right), S_{\text{global}}\left(s_q, \hat{\mathbf{v}}\right)\right)\right] \\
&\mathcal{L}_{VSM} = \lambda_1 L_{local} + \lambda_2 L_{global}
\end{aligned} \tag{3.13}$$

위 식에서 δ는 margin이고, λ_1, λ_2는 하이퍼 파라미터(hyper-parameter)이다. 각 값은 순서대로 0.1, $1e^{-2}$, 8로 설정하였다.

3.3.2 Video-QA Matching Task

본 연구에서는 제안한 모델의 추론 능력 향상을 위하여 대조 학습을 사용하는 Video-QA matching (VQAM) task라는 새로운 pre-training task를 제안한다. VQAM은 section 3.2.3에서 얻은 질의응답을 추론한 비디오 표현을 실제 질의응답과 같아지는 방향으로 학습하여 reasoning module의 추론 능력을 향상시켜주는 task이다. VQAM에서는 anchor와 positive sample, negative sample을 구성하여 대조 학습으로 진행한다. 학습 대상이 되는 anchor는 multi-modal transformer를 통해 얻은 비디오 클립과 자막의 multi-modal feature와 비디오 클립과 질의의 multi-modal feature를 attention based reasoning network를 통해 질의에 대한 답을 추론하는 멀티모달 추론 표현 H_i로 사용한다. positive sample은 해당 비디오 장면 내에서 질의와 정답을 함께 embedding module과 multi-modal transformer로 인코딩한 feature z_{pos} 를 사용한다. Negative sample은 해당 비디오 장면 내에서 질의와 오답을 함께 인코딩한 feature z_{neg} 혹은 해당 비디오 장면이 아닌곳에서의 질의와 오답을 인코딩한 feature z_{neg}를 사용하였다.

$$\begin{aligned} z_{pos} &= f_{trans}(V^{emb}, Q_{i\in\Omega}{A_{corr}}^{emb}) \\ z_{neg} &= f_{trans}(V^{emb}, Q_{i\in\Omega}{A_{incorr}}^{emb}) \\ &= f_{trans}(V^{emb}, Q_{i\notin\Omega}{A_{incorr}}^{emb}) \end{aligned} \tag{3.14}$$

해당 비디오 장면이 아닌 곳에서의 질의 응답을 negative sample로 함께 사용한 이유는 negative sample의 개수를 늘려 대조 학습를 더 효율적으로 하기 위함이다. 학습에 사용되는 샘플의 구성은 positive sample 1개에 negative sample $N-1$개로 이루어져 있고, 비디오의 scene이 바뀔 경우 새로운 샘플들로 바뀌게 된다. 학습은 인코딩된 positive QA pair z_{pos}와 멀티모달 추론 표현 H_i의 similarity를 크도록, negative QA pair z_{neg}과 H_i의 similarity는 작은 값을 가지도록 학습시킨다. 따라서 Video-QA matching task에서의 contrastive loss는 다음의 식 (3.15)와 같다.

$$\mathcal{L}_{VQM} = \sum_{i \in I} \ell_i = -\sum_{i \in I} \log \frac{\exp\left(sim\left(H_i, \boldsymbol{z}_{\text{pos}}\right)/\tau\right)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq \text{pos}]} \exp\left(sim\left(H_i, \boldsymbol{z}_k\right)/\tau\right)} \tag{3.15}$$

$\mathbb{1}$은 indicator function, N은 sample의 전체 개수, τ는 temperature parameter, cosine similarity function $sim(A, B) = A^T B / \|A\| \|B\|$이다.

3.4 비디오 질의응답 학습(Fine-tuning)

두 Video-Subtitle matching (VSM) task와 Video-QA matching (VQAM) task를 pre-training 한 후 비디오 질의응답 과제를 수행하기 위한 학습을 위해 fine-tuning을 해주는 과정이 필요하다. Fine-tuning을 위한 과정은 그림 3.3과 같다. 비디오 클립(V)과 자막(S), 그리고 비디오 클립(V)과 질의응답(QA)을 multi-modal transformer의 입력으로 넣어 각각에 대해서 통합된 표현을 얻은 후, 어텐션 기반의 reasoning network를 통해 QA-aware global representation을 만든다. 그 후로 multi-layer perceptron(MLP)과 softmax layer를 거쳐서 probability score p_{ans}를 얻어 정답을 맞출 수 있도록 학습시킨다. 학습의 목적 함수는 다음 식 (3.16)과 같다. (y_i는 i번째 질문의 정답 인덱스, N_q는 질문의 개수)

$$\mathcal{L}_{ans} = -\frac{1}{N_q} \sum_{i=1}^{N_q} \log \mathbf{p}_{ans}^{(i)} [y_i] \tag{3.16}$$

3.5 Implementation Details

본 모델은 임베딩을 얻기 위해 초당 1 프레임을 입력받고, 질의에 대한 답변을 최종 결과물로 출력한다. ImageNet 데이터셋 (Deng et al., 2009)으로 pre-trained 된 ResNet-101 (He et al., 2016)을 사용하여 1024차원의 appearance feature를 얻고, Kinetics 데이터셋 (Kay et al., 2017)으로 pre-trained 된 I3D(Carreira & Zisserman,

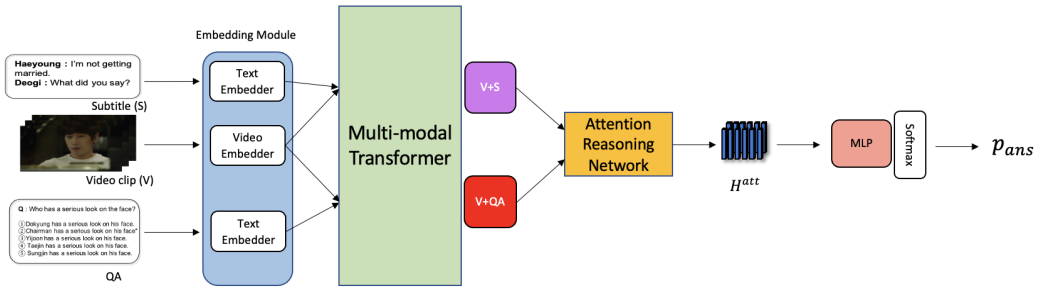

그림 3.3: 모델의 fine-tuning 구조

2017)를 사용하여 1024차원의 motion feature를 얻는다. 최종적으로 두 visual feature를 fully connected layer를 거쳐 768차원에서 concatenate 해주어 하나의 feature를 얻은 후 layer normalization을 적용한다. 언어 정보 입력(Text input)에 대해서는 WordPieces (Wu et al., 2016)로 문장들을 tokenize한 후 pre-trained 된 Bert (Devlin et al., 2018)를 이용하여 768차원의 feature를 얻은 후 layer normalization을 적용한다.

본 모델은 Pytorch (Paszke et al., 2019)로 구현되었으며, 모델을 pre-training 하기 위하여 AdamW optimizer (Loshchilov & Hutter, 2017)에 10^{-4} 값의 learning rate와 10^{-2}값의 weight decay를 사용한다. 그리고 난 후 fine-tuning을 위해 AdamW optimizer에 $5e^{-4}$의 learning rate와 $1e^{-4}$값의 weight decay를 사용하여 학습한다. VSM의 margin인 δ는 0.1으로, λ_1, λ_2는 각각 $1e^{-2}$, 8로 설정한다. Constrastive learning을 위한 QA 쌍은 N으로 표시되며 300으로 설정하여 학습한다. 전체 실험에서 하나의 mini-batch size는 4,epoch은 20, 그리고 transformer layer N_L은 3으로 설정한다. 본 모델에서 사용된 그 외의 하이퍼 파라미터(Hyper-parameter)는 표 3.1에 정리되어 있다.

	하이퍼 파라미터	설정값
Overall	Mini-batch size	4
	Epoch	20
Pre-training	Optimizer	AdamW
	Learning rate	$1e^{-4}$
	Weight decay	$1e^{-2}$
	δ	0.1
	λ_1	$1e^{-2}$
	λ_2	8
	N (# of QA Pair)	300
Fine-tuning	Optimizer	AdamW
	Learning rate	$5e^{-4}$
	Weight decay	$1e^{-4}$
	N_L	3

표 3.1: 전체 모델에 대한 하이퍼 파라미터(Hyper-parameter) 설정값

제 4 장 실험 결과 및 분석

4.1 데이터 명세

4.1.1 DramaQA

본 논문에서는 비디오 질의 응답(VideoQA) 문제의 평가를 위해 DramaQA 데이터셋 (Choi et al., 2020)을 사용하였다. DramaQA 데이터셋은 한국 드라마인 "*또 오해영*"에서 데이터를 수집해 총 18개의 에피소드에서 총 20.5시간의 길이를 가진 $23,928$개의 비디오 클립으로 구성되어 있다. 각 비디오 클립은 다양한 길이를 가지며 비디오 프레임은 초당 3 프레임으로 구성되며, 네 가지 어려움 단계로 구성된 $17,983$개의 드라마에 대한 질의 응답을 포함하고 있다. 학습과 검증, 그리고 평가를 위해 이 데이터셋은 각각 $11,118$개와 $3,412$개, $3,453$개로 나누어져 있다. 또한 이 데이터셋은 시각적 bounding box과 주요 인물들의 행동 및 감정, 이름이 tagging 된 대본 등에 대한 문자 중심 annotation들이 포함되어 있다.

DramaQA 데이터셋의 질의 응답 쌍은 어려움의 정도를 memory capacity와 logical complexity 기준에 따라 네 가지 레벨로 나눈다. Memory capacity는 질의에 답하기 위해 필요한 비디오의 요구되는 길이를 의미하며, 비디오의 길이가 길수록 문제에 대한 답변을 위한 추론(Reasoning)이 어려워지게 된다. 그러므로 본 데이터셋은 memory capacity를 기반으로 비디오 클립의 종류에 따라 레벨이 나누었다.

i. **Level 1 (Shot):** 비디오가 10초보다 짧은 길이를 가지며 단일 카메라 앵글에서의 shot으로 이루어진다. 대부분의 VideoQA datasets (Lei et al., 2018; Tapaswi et al., 2016; K.-M. Kim et al., 2017)이 이 단계의 어려움을 가진다.

ii. **Level 2 (Scene):** 비디오가 장소의 변동없이 $1-10$분 정도의 길이로 구성되

어 있고, 1 단계의 shot들이 모인 일련의 행동들로 구성되어 있다. 이 단계를 *"story" 단계*라고도 하며 질문에 대한 답변을 위해 더 어려운 수준의 추론을 요구한다.

이 두 단계의 레벨에서 logical complexity를 기준으로 다시 네 단계로 나눈다. Logical complexity는 질문에 답하기 위해 얼마나 많은 logical step이 필요한가에 대한 것으로 인과 관계가 필요하거나 짧은 장면에서 다양한 단서를 찾아야하는 질의가 나올수록 어려움 레벨이 높아진다. 또한 추가적인 시각 정보로 비디오 내 등장인물들의 bounding box와 행동 및 감정 정보가 함께 포함된다. 그래서 최종적으로 다음과 같은 네 가지의 단계로 나누어지고, 본 논문에서는 이 네 가지 단계에 대해 다른 모델들과 비교하여 평가를 진행한다.

i. **Diff 1 :** 이 단계의 질의 응답 쌍은 Level 1의 난이도를 가지며, 동영상으로부터 얻은 하나의 사실을 기반으로 구성되어 있다. 10초의 짧은 길이를 가진 비디오에 해당하며, 이 단계에서는 생각들을 서로 분리하고 결합시키는 인지적인 추론을 하지 않고 직접적으로 보이는 사람과 하나의 물체간의 관계와 같은 것을 찾는 단계이다. 한 사실은 *subject-relationship-object*와 같은 형태의 triplets로 표현된다. 각 질문은 *Who, Where,* 그리고 *What*으로 시작된다.

ii. **Diff 2 :** 이 단계의 질의 응답 쌍은 Level 1의 난이도를 가지며, 동영상으로부터 얻은 여러개의 사실을 기반으로 구성되어 있다. 1분 내의 길이를 가진 비디오에 해당하고, 사람과 여러 사람 혹은 여러 물체와의 관계 및 사실을 이용하여 간단한 추론을 요구하는 단계이다. 각 질문은 *Who, Where,* 그리고 *What*으로 시작된다.

iii. **Diff 3:** 이 단계의 질의 응답 쌍은 Level 2의 난이도를 가지며, 일련의 정보와 함께 여러개의 상황과 행동들을 기반으로 구성되어 있다. Diff 2와 다르게 상황이 어떻게 변하고 있는지 등을 파악하기 위해 다중 정보들을 결합하여 질

문에 대한 정답을 찾아야한다. 1 – 3분정도의 길이를 가진 비디오를 사용하고, 답변하기 위해 여러 순간의 supporting facts들을 사용하여 주인공의 상황이 어떻게 바뀌었는지, 그에 따라 어떻게 행동했는지에 대해 질문을 하게 된다. 각 질문은 *How*와 *What*으로 시작된다.

iv. **Diff 4:** 이 단계의 질의 응답 쌍은 Level 2의 난이도를 가지며, 질문이 상황이나 행동에 대한 인과 관계의 추론을 기반으로 구성되어 있다. 여기서는 1 – 10분의 긴 길이의 비디오를 사용하게 된다. 이 수준의 질문에서는 ”Why”로 시작할 수 있는 인과관계에 대한 추론을 다루게 된다. 인과관계 추론은 인과관계를 확인하는 과정이며, 사람의 행동이나 상황의 원인과 결과 사이의 관계를 추론하는 것이므로 앞의 단계보다 복잡하고 어려운 질문이라고 할 수 있다. 각 질문은 *Why*로 시작된다.

4.2 평가 지표

평가에 사용될 DramaQA 데이터셋은 네 가지 어려움 단계로 나누어지고, 각각의 어려움이 사람 수준의 인지 발달 능력을 평가할 수 있도록 구성되어 있다. 정량적인 평가 지표는 평가 질의에 대한 정답률(%)로 나타내며, 각 난이도 마다 정답률을 측정하고 전체 질의에 대한 평균 정답률과 함께 평가한다. ECCV 2020 DramaQA 챌린지에 나온 모델들과 논문을 기반으로 한 최신 모델들과 성능을 비교한다.

4.3 정량 평가

4.3.1 다른 연구와의 비교

본 연구에서 제안한 모델을 평가하기 위해 ECCV 2020에서 열린 DramaQA challenge의 상위 5개 팀과의 성능을 비교하였고, 그 결과는 표 4.1과 같다. 평가 기준은 DramaQA challenge의 기준을 따르며, 평가는 네 가지 어려움 단계의 각각에 대한 평

가와 전반적인 결과로 구성된다. 표 4.1의 결과를 보면 모든 난이도에서 다른 모델과 비교했을 때 가장 좋은 성능을 보였고, Overall rate은 가장 높은 수치를 가지는 것을 확인할 수 있었다. Diff 4의 경우 제일 높은 성능인 GGANG 모델과 비교했을 때 1점 정도의 차이가 나고, Diff 3의 경우에는 나머지 다섯 모델과 비교해 8% 이상의 압도적인 성능 향상을 보였음을 확인할 수 있었다. 이는 본 연구에서 제안한 모델이 어려운 수준의 비디오 질의응답 문제를 효과적으로 해결할 수 있음을 입증한다.

또한 기존에 비디오 질의응답 문제를 해결하기 위해 제안한 모델들과의 비교 실험도 진행하였고, 그 결과는 표 4.2와 같다. 본 실험은 DramaQA 데이터셋에서 test set이 공개되어 있지않아 validation set을 기준으로 평가하였다. DramaQA 논문이 제안한 모델과 비교했을 때 본 논문에서 제안한 모델이 어려운 질의에서 큰 성능 향상을 보였음을 확인할 수 있었고, 마찬가지로 Diff 3와 Diff 4에서 가장 좋은 성능을 보였다. 위 결과에서 주목할만한 점은 질의의 어려움이 올라갈수록 성능 하락폭이 다른 최신 모델에 비해 크지 않다는 점이다. 이것은 본 모델이 질의의 어려움 정도에 크게 영향을 받지 않고 강인한 성능을 낼 수 있음을 시사한다.

Challenger	Diff 1	Diff 2	Diff 3	Diff 4	Overall
IITDrama	76	72	55	60	71
bjorn	77	74	57	57	71
HARD KAERI	76	73	56	59	71
Sudoku	78	74	68	67	75
GGANG	81	79	64	70	77
Ours	**83**	**82**	**73**	**71**	**79**

표 4.1: ECCV 2020의 DramaQA challenge 결과.

Models	Diff 1	Diff 2	Diff 3	Diff 4	Overall
DramaQA (Choi et al., 2020)	76.0	74.7	57.4	56.6	71.1
(S. Kim et al., 2020)	**84.0**	**85.0**	70.0	70.0	**81.0**
(Bebensee & Zhang, 2021)	80.6	78.4	68.5	68.7	77.2
Ours	83.1	81.9	**73.3**	**71.2**	79.2

표 4.2: DramaQA 데이터셋에 대한 기존 비디오 질의응답 모델들과의 비교 평가 결과.

4.3.2 Ablation Study

우리는 본 장을 통해 제안한 모델의 각 모듈의 효과와 pre-training task의 유효성을 파악하고, 하이퍼 파라미터 셋팅에 따른 결과도 함께 분석한다.

표 4.3는 본 모델에서 핵심 모듈의 유무에 따른 실험 결과를 기재해놓은 것이다. Multi-modal Transformer와 어텐션 기반 reasoning network를 모두 사용하지 않고 supervised learning으로 실험한 dot-product 결과와 MLP 모델의 결과를 보면, 전반적인 난이도에서 모두 낮은 성능을 보이는 것을 통해 비디오의 표현을 충분히 학습하지 못했음을 확인할 수 있다. Multi-modal fusion 방법으로 multi-modal encoder 대신 late fusion을 한 LSTM (Hochreiter & Schmidhuber, 1997) 모델로 실험을 해본 결과(Ours - MulEN)와 비교했을 때, 짧은 비디오에 대한 질의응답인 Diff 1과 Diff 2에서는 비교적으로 좋은 성능을 보이지만, 비디오 scene 단위로 구성된 Diff 3과 Diff 4에서는 long-term dependency를 해결하지 못하여 좋지 않은 성능을 보였음을 알 수 있다. 이런 결과를 통해 우리가 제안한 multi-modal transformer가 long-term dependency를 효과적으로 활용하였음을 입증하였다. 비디오 표현과 질의를 attention을 이용하지 않고 두 표현을 concatenate 하여 MLP를 이용한 경우에는 (Ours - Attention), 비디오에서 쉽게 찾을 수 있거나 한 개의 단서만으로 문제를 풀 수 있는

Diff 1에서는 성능 하락이 눈에 띄진 않는다. 하지만 추론 능력이 필요하거나 비디오에서 짧은 시간내에 많은 단서를 찾아내야 하는 Diff 2, 3, 4 질의에서는 attention을 사용했을 때보다 큰 성능 하락을 볼 수 있어, 우리의 모델에서 사용한 어텐션 기반 reasoning module이 어려운 질의에 더 잘 대처하도록 학습된 것을 알 수 있었다.

Model	Diff 1	Diff 2	Diff 3	Diff 4	Overall
Dot-product (Q+A)	30.3	27.7	28.2	27.4	28.9
MLP (Q+A)	51.5	47.7	41.5	50.4	50
Ours - MulTr	57.3	61.1	48.1	49.2	54.4
Ours - Attention	75.3	67.2	59.0	51.4	65.4
Ours	**83.1**	**81.9**	**73.3**	**71.2**	**79.2**

표 4.3: 핵심 Module의 유무에 따른 실험

표 4.4는 모델의 구조는 그대로지만 pre-training을 하지 않고 supervised setting으로 학습하였을 때와 pre-training setting으로 학습했을 때의 결과를 비교하여 보여준다. 모든 난이도에 대해 전반적으로 성능 차이가 나는 것으로 보아, pre-training task가 추가함으로써 multi-modal transformer가 갖는 비디오의 표현력 및 장면 이해도와 추론 능력을 향상시켰다는 것을 알 수 있다.

Model	Diff 1	Diff 2	Diff 3	Diff 4	Overall
w/o pre-training	73.4	72.6	62.2	59.7	67.4
Ours	**83.1**	**81.9**	**73.3**	**71.2**	**79.2**

표 4.4: Pre-training 유무에 따른 실험 결과.

표 4.5에서는 계층 구조를 가지고 있는 multi-modal transformer에서 계층 수에 따

른 정답률을 기재하였다. Level 1에서는 cross-modal transformer를 이용하여 local한 feature들을 뽑았고, local feature들을 LSTM의 입력으로 넣어준 후 concatenate 하는 late-fusion 방법을 이용하여 실험을 하였다. 여기서는 local feature들을 global context로 합성시켜주어 long-term dependency를 강하게 해주는 temporal transformer를 사용하지 않아서, 상대적으로 긴 비디오에 대해 질의응답인 diff $3, 4$에서 많은 성능 하락이 있음을 볼 수 있었고, 추가적으로 전반적인 비디오 표현력이 떨어져서 diff $1, 2$에서도 성능이 많이 떨어지는 것을 알 수 있었다. Level 2에서는 cross-modal transformer와 temporal transformer 두 개의 transformer를 두 층으로 쌓아 multi-modal fusion을 해준 결과이다. 마지막 층에 cross-modal transformer를 적용하기 전과 비교했을 때, Diff $1, 2$에서는 비슷하거나 더 높은 성능이 나오지만 Diff $3, 4$에서는 큰 성능 차이가 나는 것으로 보아 global context 상에서 cross-attention을 해주었을 때에 long-term dependency가 더 효과적으로 인코딩되어 활용하는 것을 알 수 있었다.

Model	Diff 1	Diff 2	Diff 3	Diff 4	Overall
Level 1 (Cross-modal Transformer)	71.2	70.9	58.3	50.2	64.5
Level 2 (Cross-modal + Temporal Transformer)	80.6	80.5	69.3	65.4	75.7
Ours	**83.1**	**81.9**	**73.3**	**71.2**	**79.2**

표 4.5: Hierarchical Multi-modal transformer의 층 수에 따른 실험 결과.

그림 4.1은 대조학습을 이용한 Video-QA Matching task에서 사용된 negative sample의 개수에 따른 실험 결과이다. 일반적으로 대조 학습에서는 negative sample의 개수가 많을수록, 그리고 negative sample이 hard negative인 sample일수록 더 좋은 학습을 할 수 있다. 그래프의 결과를 통해 negative sample의 개수를 적게 사용하여 학습을 하면 pre-training이 효과를 가지지 못하는 것과 같은 성능을 보이며,

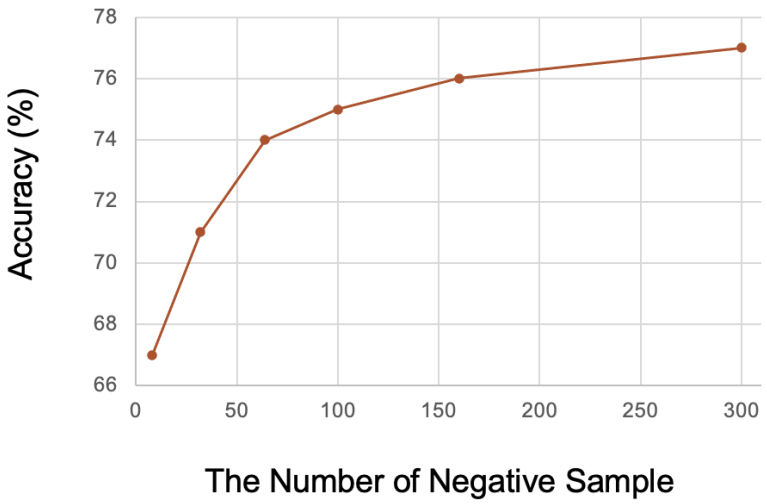

그림 4.1: Video-QA Matching (VQAM) task에서 negative sample의 개수에 따른 실험 결과.

negative sample의 개수를 늘려서 실험하면 처음에는 가파르게 성능이 올라가다가 어느 지점부터는 거의 성능이 올라가지 않는 것을 확인할 수 있었다. 그래프에서는 negative sample의 개수를 적게 사용하여 학습을 하면 사실상 해당 pre-training이 제대로 되지 않는 것과 같은 성능을 보이며, negative sample의 개수를 늘려서 실험하면 처음에는 가파르게 성능이 올라가다가, 약 150개의 negative sample을 가지는 지점부터는 성능이 거의 올라가지 않는 것을 알 수 있다.

제 5 장 결론

비디오 질의응답 문제는 시각 정보와 언어 정보를 같이 활용하여 인공지능 모델을 인간의 인지 과정을 모사할 수 있도록 발전시키는 데에 큰 도움을 주고 있어 매우 중요한 과제 중 하나이지만, 비디오의 길이가 길어지고 복잡한 추론 능력을 요구하는 경우 장기적인 의존성 고려와 추론 과정에서 어려움을 겪고 있다.

본 논문에서는 기존의 짧은 길이의 비디오 클립에서의 질의응답이 아닌 복잡하고 긴 비디오 클립 내의 장기 의존성 문제를 해결하면서 어려운 질의응답 문제를 위한 추론 능력을 향상시키기 위해 계층적 구조를 가진 새로운 멀티모달(Multi-modal) transformer 구조와 대조 학습(Contrastive learning)을 이용한 새로운 pre-training task를 제안하였다. 짧은 지역적인 비디오 클립 여러 개를 하나로 통합시켜 주는 temporal transformer와 언어적 정보와 시각적 정보를 매칭시켜 멀티모달 표현을 얻게 해주는 cross-modal transformer로 이루어진 세 단계의 계층 구조로 이루어진 multi-modal transformer를 제안하여 장기 의존성을 인코딩하고 비디오 장면에 대한 이해력을 향상시켰다. 먼저 cross-modal transformer를 이용하여 짧은 비디오 클립에서의 비디오 프레임들과 자막을 local context 관점에서 정렬시켜 하나의 비디오 표현을 얻게 하고, 그 후 temporal transformer를 이용하여 앞서 얻은 여러 시간대의 cross-modal feature들을 하나로 합쳐 비디오의 전체적인 맥락을 이해할 수 있는 표현을 얻게 해주고 마지막으로 global context에서 언어 정보와 시각 정보를 교환하고 정렬하여 맞춰줌으로써 최종적으로 비디오 표현을 얻게 된다. 비디오 질의응답을 위한 모델을 효과적으로 pre-training 하기 위해, 앞서 multi-modal transformer에서 얻은 비디오 표현을 사용하고, Video-Subtitle Matching task와 새롭게 제안한 Video-QA Matching task를 적용함으로써 질의응답을 위한 추론 능력을 향상시킨다. VSM

task에서는 비디오 입력의 자막과 영상간의 매칭을 학습시켜 local alignment을 얻게 해주고, 비디오의 자막이 다른 비디오에 해당하지 않도록 학습시켜 global alignment를 얻게 해주어 multi-modal transformer가 더 효과적인 비디오 표현을 얻을 수 있게 도와준다. VQAM task에서는 multi-modal transformer에서 얻은 비디오 표현과 질의응답 간의 대조 학습을 통해, 비디오 표현과 질의응답의 정답간의 유사도를 크게 학습하고 비디오 표현과 오답간의 유사도를 낮도록 학습하여 비디오 표현이 질의응답에 대해 정답과 가깝게 추론할 수 있도록 만든다. 실험으로는 네 가지 어려움 단계로 구분된 DramaQA 데이터셋에서 평가를 진행하였다. ECCV 2020 DramaQA challenge와의 결과 비교로 모든 난이도에서 가장 높은 성능을 보였으며, 최신 논문들과의 비교로는 $1-10$분 사이의 긴 비디오에서의 복잡한 질의응답인 Diff 3,4에서 각각 3%,1% 이상의 차이를 보여 본 모델이 긴 비디오와 복잡한 질의응답을 잘 해결할 수 있다는 것을 증명할 수 있었다. 또한 ablation study로 본 모델에서 사용한 핵심 모듈들과 pre-training task의 유무에 따른 성능 향상을 보여 유효성을 검증하였다.

본 연구에서는 다양한 비디오 태스크와 데이터셋 중에서 비디오 질의응답 문제, DramaQA라는 하나의 데이터셋에서만 실험되었다. 따라서 추후 연구로는 VideoQA의 다양한 데이터셋에서의 실험과 다른 모델과의 비교가 추가로 필요할 것이고, pre-training을 통해 학습된 추론에 강점을 가진 모델인 만큼 다양한 Vision-language 태스크 (e.g. Video-Retrieval, Video-language inference, Video Captioning, Open-ended VideoQA)에 대해 추가로 수행하여 모델의 pre-training 성능을 측정할 것이다. 그리고 현재는 video-subtitle matching, video-QA matching 두 가지의 pre-training 방법만을 채택하여 사용하고 있지만, 질문에 해당하는 비디오의 구간을 찾아내거나 비디오의 순서를 뒤바꾼 후 올바른 순서를 찾게하는 방법과 같이 비디오의 표현력을 강화시킬 pre-training task를 새롭게 구성하여 추가적인 학습을 하는 것을 추후 연구방향으로 두고 있다.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the ieee international conference on computer vision* (pp. 2425–2433).

Bebensee, B., & Zhang, B.-T. (2021). Co-attentional transformers for story-based video understanding. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4005–4009).

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6299–6308).

Carreira-Perpinan, M. A., & Hinton, G. E. (2005). On contrastive divergence learning. In *Aistats* (Vol. 10, pp. 33–40).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Choi, S., On, K.-W., Heo, Y.-J., Seo, A., Jang, Y., Lee, S., … Zhang, B.-T. (2020). Dramaqa: Character-centered video story understanding with hierarchical qa. *arXiv preprint arXiv:2005.03356*.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92–103.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer*

vision and pattern recognition (pp. 248–255).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2015). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, *38*(9), 1734–1747.

Gao, J., Ge, R., Chen, K., & Nevatia, R. (2018). Motion-appearance co-memory networks for video question answering. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6576–6585).

Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9729–9738).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., ... Sumi, K. (2017). Attention-based multimodal fusion for video description. In *Proceedings of the ieee international conference on computer vision* (pp. 4193–4202).

Jang, Y., Song, Y., Yu, Y., Kim, Y., & Kim, G. (2017). Tgif-qa: Toward spatio-temporal

reasoning in visual question answering. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2758–2766).

Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, *35*(1), 221–231.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kim, K.-M., Heo, M.-O., Choi, S.-H., & Zhang, B.-T. (2017). Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*.

Kim, S., Jeong, S., Kim, E., Kang, I., & Kwak, N. (2020). Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. *arXiv preprint arXiv:2009.08043*.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Carlos Niebles, J. (2017). Dense-captioning events in videos. In *Proceedings of the ieee international conference on computer vision* (pp. 706–715).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Lei, J., Yu, L., Bansal, M., & Berg, T. L. (2018). Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Lei, J., Yu, L., Berg, T. L., & Bansal, M. (2020). Tvr: A large-scale dataset for video-subtitle moment retrieval. *arXiv preprint arXiv:2001.09099*.

Li, L., Chen, Y.-C., Cheng, Y., Gan, Z., Yu, L., & Liu, J. (2020). Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Icml*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*, 8026–8037.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2536–2544).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Snoek, C. G., Worring, M., & Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual acm international conference on multimedia* (pp. 399–402).

Socher, R., Chen, D., Manning, C. D., & Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems* (pp. 926–934).

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016).

Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4631–4640).

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. association for computational linguistics. meeting* (Vol. 2019, p. 6558).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*(12).

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, *abs/1609.08144*. Retrieved from `http://arxiv.org/abs/1609.08144`

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., & Zhuang, Y. (2017). Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th acm international conference on multimedia* (pp. 1645–1653).

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5288–5296).

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *Acm computing*

surveys (CSUR), *38*(4), 13–es.

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Zeng, K.-H., Chen, T.-H., Chuang, C.-Y., Liao, Y.-H., Niebles, J. C., & Sun, M. (2017). Leveraging video descriptions to learn video question answering. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision* (pp. 649–666).

Zhang, R., Isola, P., & Efros, A. A. (2017). Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1058–1067).

Abstract

Video Question Answering (VideoQA) task requires the ability to effectively understand various relationships in the real-world using visual and linguistic information of video. It is an effective task to evaluate the ability of AI agents to incorporate human-level intelligence. However, the previous videoQA methods conducted in short-length video clips. Question-and-answer of long complex videos as well as short-length videos must be addressed using long-term dependency and high-level reasoning capabilities. In this paper, we propose a novel multi-modal transformer with a hierarchical structure and pre-training task using contrastive learning to improve reasoning ability for difficult VideoQA problems while solving the long-term dependency problem within complex and long video clips. Proposed multi-modal transformer with a three-level hierarchical structure to encode long-term dependencies improves the understanding of video scenes and to learn the context of the video and to infer the correct answer by using the video representation that is highly related to the query. In addition, the Video-Subtle Matching (VSM) task and the newly proposed Video-QA Matching (VQAM) task are used to effectively pre-train the correct answer to complex questions as well as to learn the representation of the video. In particular, VQAM, newly proposed in this paper, significantly enhances reasoning ability for question-answering by using attention and contrast learning. For the evaluation of our framework, we conduct experiments using the Korean DramaQA dataset with four difficulty levels and show the state-of-the-art performance compared to previous methods. In addition, we demonstrate the effectiveness of each proposed module through an ablation study.

Keywords: Deep learning, VideoQA, Multi-modal learning, Pre-training

Student Number: 2018-22635