The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.3, pp.140~147

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.3.140

Three-Dimensional Convolutional Vision Transformer for Sign Language Translation

Horyeor Seong† · Hyeonjoong Cho††

ABSTRACT

In the Republic of Korea, people with hearing impairments are the second-largest demographic within the registered disability community, following those with physical disabilities. Despite this demographic significance, research on sign language translation technology is limited due to several reasons including the limited market size and the lack of adequately annotated datasets. Despite the difficulties, a few researchers continue to improve the performacne of sign language translation technologies by employing the recent advance of deep learning, for example, the transformer architecture, as the transformer-based models have demonstrated noteworthy performance in tasks such as action recognition and video classification. This study focuses on enhancing the recognition performance of sign language translation by combining transformers with 3D-CNN. Through experimental evaluations using the PHOENIX-Wether-2014T dataset [1], we show that the proposed model exhibits comparable performance to existing models in terms of Floating Point Operations Per Second (FLOPs).

Keywords : Sign Language Translation, Transformer, Convolutional Transformer

수어 번역을 위한 3차원 컨볼루션 비전 트랜스포머

성 호 렬† · 조 현 중††

요 약

한국에서 청각장애인은 지체장애인에 이어 두 번째로 많은 등록 장애인 그룹이다. 하지만 수어 기계 번역은 시장 성장성이 작고, 엄밀하게 주석처리가 된 데이터 세트가 부족해 발전 속도가 더디다. 한편, 최근 컴퓨터 비전과 패턴 인식 분야에서 트랜스포머를 사용한 모델이 많이 제안되고 있는데, 트랜스포머를 이용한 모델은 동작 인식, 비디오 분류 등의 분야에서 높은 성능을 보여오고 있다. 이에 따라 수어 기계 번역 분야에서도 트랜스포머를 도입하여 성능을 개선하려는 시도들이 제안되고 있다. 본 논문에서는 수어 번역을 위한 인식 부분을 트랜스포머와 3D-CNN을 융합한 3D-CvT를 제안한다. 또, PHOENIX-Wether-2014T [1]를 이용한 실험을 통해 제안 모델은 기존 모델보다 적은 연산량으로도 비슷한 번역 성능을 보이는 효율적인 모델임을 실험적으로 증명하였다.

키워드 : 수어 번역, 트랜스포머, 컨볼루션 트랜스포머

1. 서 론

수어는 말 대신 수지 요소(손동작)와 비수지 요소(표정, 몸짓, 등)를 사용하는 시각 언어로 청각장애인과 의사소통하는데 사용되며, 음성 언어를 단순히 시각적으로 표현한 것이 아닌, 고유한 자체 문법, 구문, 어휘를 가진 언어이다.

2021년 세계보건기구는 전 세계적으로 4억 3천만 명 이상의 청각장애인이 장애를 겪고 있다고 추정했는데, 이는 전 세계 인구의 5.4%에 해당하는 수치이다. 청각장애인들은 청인들에 비해 다양한 정보에 대한 접근성이 현저히 떨어지므로, 이를 개선하기 위한 수어 기계 번역에 대한 관심이 증대되고 있다.

수어 번역은 시각 언어인 수어를 청각 언어인 음성 언어로 번역하는 과정이다. 수어와 음성 언어는 동일한 지역에서 사용된다 하더라도 별도의 언어 체계이므로 수어 번역은 시각 언어를 청각 언어로 번역해야하는 어려운 작업이다.

최근 딥러닝의 발달과 더불어 다양한 수어 기계 번역 모델이 소개되었다[1-6]. 일반적으로 수어 번역 모델은 비전 특징 추출 단계과 번역 단계의 두 단계로 구성된다.

첫 번째 단계로 수어 비디오를 비전 특징 추출기(Vision

※ 본 연구는 2023년 정부(교육부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. NRF-2021R1F1A1049202).
† 준 회 원 : 고려대학교 컴퓨터정보학과 석사
†† 종신회원 : 고려대학교 컴퓨터융합소프트웨어학과 교수
Manuscript Received : November 22, 2023
First Revision : January 23, 2024
Accepted : February 9, 2024
* Corresponding Author : Hyeonjoong Cho(raycho@korea.ac.kr)

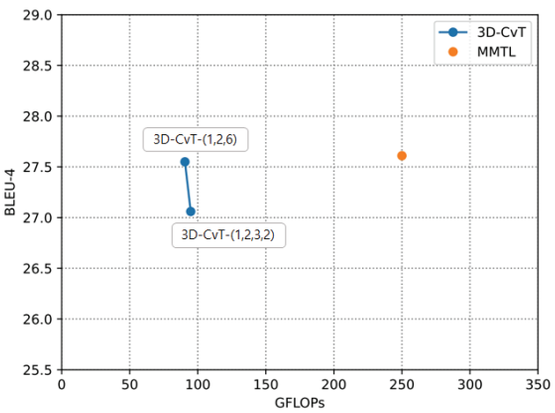


Fig. 1. Translation Performance(BLEU) over Computation (GFLOPs) between MMTL(SoTA) and 3D-CvT(Ours)

Feature Extractor, VFE)에 입력하여 시각적 특징을 추출한다. 그 후 분류기가 해당 수어의 수어소(Gloss)를 분류하는데, 수어소는 최소 의미 단위의 수어에 대한 주석을 의미한다. VFE로 3D-CNN, RNN, LSTM 등의 모델이 빈번히 사용된다.

두 번째 단계에서는 추출된 수어 특징 또는 수어소를 번역 모델의 입력으로 활용하여 청각 언어를 생성한다. 수어 번역은 수지 요소와 비수지 요소 입력을 사용하기 때문에 컴퓨터 비전 분야이기도 하지만, 수어에서 청각 언어로 번역하는 기계 번역의 한 분야이기도 하다. 이러한 이유로 번역 모델로 NMT(Neural Machine Translation)에 사용되는 모델을 적극적으로 도입하려는 시도들이 있어 왔다.

상기 기술한 첫 번째 단계인 VFE에는 CNN 기반 방법과 트랜스포머 기반 방법이 있다. CNN은 여러 단계의 특징 맵을 생성하여 각 특징 맵에서 지역적인 특징을 효과적으로 추출할 수 있다. CNN은 학습 속도가 빠르고 로컬 정보를 잘 학습한다는 이점이 있다. 반면 비전 트랜스포머는 입력 신호를 패치 단위로 나누어 어텐션(Attention)을 통해 특징을 추출하는데, CNN에 비해 전체적인 전역 정보를 파악하는 데 효과적이다. 한편, 이들 두 기술의 장점을 취하고자 CNN과 트랜스포머를 융합한 CvT(Convolutional vision Tansformer) [8]가 최근 소개되기도 했다.

비전 트랜스포머(ViT) [7]의 출시 이후, 대부분의 컴퓨터 비전 영역에서 ViT 기반 모델이 등장했고, 기존의 CNN 기반 모델과 성능 면에서 경쟁을 하고 있다. 특히, 비디오 데이터를 위한 모델의 경우, 현재 SoTA(State of The Art) 모델은 CNN [15,16]이 아닌 ViT를 백본으로 사용한다. 그러나 최근까지도 수어 인식 영역에서는 CNN 기반 VFE를 주로 사용해 왔다 [2,5,6].

트랜스포머를 수어 인식과 번역에 활용하기 위해 추가적으로 고려할 사항은 시간적으로 변하는 수어 동작의 특징을 추출해야 한다는 점이다. 하지만, ViT, CvT 등은 이미지 처리 모델로 설계되어 시간축을 따라 변하는 동적인 정보를 추출하는 데 한계가 있다.

본 논문에서는 CvT를 시간 축으로 확장하여 수어 영상의 특징을 효과적으로 추출하기 위한 3D-CvT를 제안한다. 또, 3D-CvT가 데이터셋 PHOENIX-2014T [1]에 대해 더 적은 연산량(FLOPs)으로도 기존 SoTA 모델과 유사한 성능을 보이는 효율적인 모델임을 실험적으로 증명한다. 실험 결과는 그림.1로 요약할 수 있는데, 제안 모델이 기존 SoTA 모델(MMTL)에 비해 적은 연산량(X축)으로도 기존과 비슷한 번역 성능(Y축)을 보임을 알 수 있다.

본 논문은 다음과 같이 구성되었다. 2장에서는 수어와 관련된 연구 현황을 살펴본다. 3장에서는 제안 기술, 3D-CvT를 구체적으로 설명한다. 그리고, 4장에서는 실험을 통해 3D-CvT의 성능을 기존 기술과 비교한다. 마지막으로 5장에서 본 논문의 결론을 맺는다.

2. 관련 연구

2.1 수어 번역 모델

수어 모델는 일반적으로 Fig. 2와 같이 비전 특징 추출기(VFE)와 번역 모델, 두 가지 구성 요소로 나눌 수 있다.

먼저 VFE는 수어 비디오 입력으로부터 수어의 특징을 추출하고, 분류기는 추출된 수어 특징을 기반으로 수어소를 생성한다. 수어 비디오가 수어 문장으로 구성된 비디오인 경우 Connectionist Temporal Classification(CTC Loss) [9]을 사

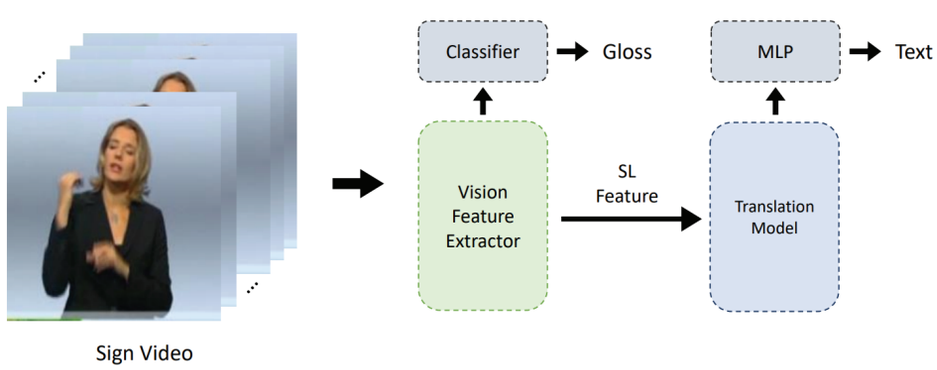


Fig. 2. The Common Structure of Sign Language Translation Models

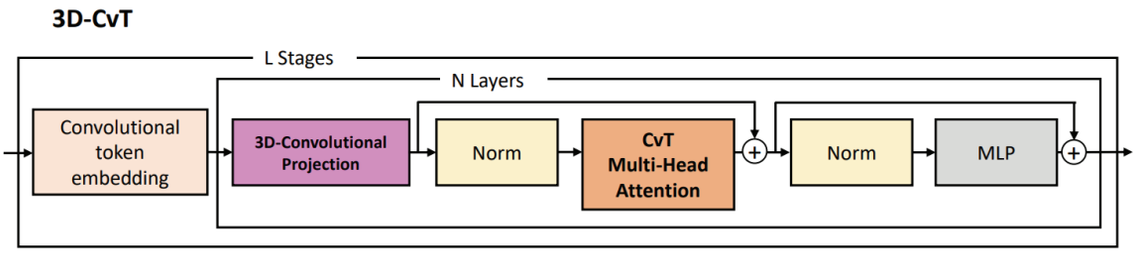


Fig. 3. The Structure of 3D-CvT

용하여 수어소의 배열을 추출할 수 있다. VFE가 시각 분야라면 번역 모델은 자연어 처리(NLP)의 신경망 기계 번역 (NMT) 분야이다. 번역 모델에서는 RNN, CNN, 어텐션(Attention) 모델과 같은 NMT용 모델을 도입하여 사용해 왔다. 수어 번역 모델에서 트랜스포머를 사용하기 시작한 것은 Camgoz 등 (2020)이 SLT [2]를 발표한 이후부터다.

2.2 수어 번역 사전 학습 및 전이 학습

사전 학습(Pretraining)과 전이 학습(Transfer learning)은 머신러닝과 딥러닝에서 중요한 두 가지 기술이며 다양한 작업에서 모델의 성능을 향상시키는 데 사용할 수 있다. Chen 등 (2022)이 소개한 MMTL (Multi-Modality Transfer Learning) [5] 이전에 수어 번역의 성능은 수어 데이터셋의 부족으로 인해 NMT에 비해 현저히 떨어졌다. 반면 MMTL은 점진적인 사전 학습과 전이 학습을 통해 수어 번역의 성능을 크게 향상했다.

최근에는 번역 모델로 다국어 번역 모델인 mBART [13]를 활용하는 경향이 있다. mBART (multilingual Bidirectional and Auto-Regressive Transformer)는 BART [19]를 다국어 번역 목적으로 학습시킨 모델이다. MMTL은 CC25 말뭉치로 사전 훈련된 mBART를 PHOENIX2014T 수어 데이터셋를 사용하여 Gloss2Text로 학습했다. 마지막으로 VFE의 파라미터를 고정하고 Sign2Text를 종단간 전체적으로 훈련시킨다. 이때 손실 함수(Loss)는 Sign2Gloss의 CTC (Connectionist Temporal Classification) 손실함수와 Sign2Text의 Cross-entropy 손실함수를 결합하여 사용한다.

2.3 컨볼루션 비전 트랜스포머

CNN이 주류이던 컴퓨터 비전 작업에 ViT가 등장한 이후 이미지 인식, 분류 등을 위한 모델 대부분이 ViT를 백본으로 사용하기 시작했다. 한편, CNN과 ViT는 모두 고유한 장단점이 있는데, 컨볼루션 비전 트랜스포머(CvT)는 CNN의 장점(예: Shift-Invariance)을 ViT 구조에 도입하는 동시에 트랜스포머의 장점(예: Global correlation awareness)을 유지한다. CvT는 ViT와 같이 입력 이미지를 패치로 분할하여 패치 임베딩을 생성한 다음, 컨볼루션 투영을 통해 쿼리, 키, 값을 생성한 후 어텐션을 진행한다. 분할된 패치에 대해서만 연산하는 ViT와 달리 CvT는 컨볼루션 연산을 적용하여 효과적으로 로컬 정보를 얻을 수 있으며, 기존 ViT보다 적은 파라미터와 적은 연산량(FLOPs)으로도 좋은 성능을 보이기도 한다.

3. 제안 방법

3.1 Why CvT?

비디오 트랜스포머 [15-18]는 기본적으로 ViT 기반 모델이다. ViT는 각 프레임을 특정 수의 패치로 잘라서 사용되고, 어텐션 연산을 통해 모든 패치들의 상관 관계를 파악한다. Fig. 4는 동작 인식 데이터셋인 Kinetics400 [11]과 수화 데이터셋인 PHOENIX2014T [1]의 프레임의 예시이다. Kinetics400의 경우 '축구공 차기'라는 동작을 인식하기 위해 공, 사람의 동작, 전체 배경에 대한 정보가 필요하지만, PHOENIX2014T의 경우 수화자의 얼굴, 손, 팔의 위치와 같은 로컬 정보가 상대적으로 더 중요하게 작용한다. 따라서 수어 영상의 특징을 효과적으로 추출하기 위해, 패치 단위로 나누어 동작하는 ViT보다, 지역 정보를 효과적으로 획득할 수 있는 컨볼루션 연산을 ViT에 도입한 CvT를 사용하는 것이 유리하다.

ViT는 입력 이미지를 패치로 나누고 평탄화(Flatten)한 각 패치를 사용하여 쿼리, 키, 밸류를 생성하지만 CvT는 패치를 다시 입력 데이터로 재구성하고 컨볼루션을 통해 쿼리, 키, 밸류를 생성한다는 점에서 차이가 있다. CvT와 유사한 구조를 가진 모델은 CMT (Convolution Meets Transformer)[20]인데, CMT 역시 이미지를 패치로 나누고 컨볼루션을 사용한다. CMT는 CvT와 유사한 모델 구조로 되어 있지만, 파라미터와 연산량를 줄이기 위해 적극적으로 경량화된 모델이다. 하지

Fig. 4. Sample Frames of PHOENIX2014T (left) and Kinetics400 (right).

만, 이미지 처리를 위해 경량화된 CMT 모델을 비디오 데이터에 직접적으로 적용할 경우 시간적인 맥락 정보를 가진 비디오에 대한 인식 성능 저하를 유발할 수 있어 여기서는 CvT를 기본 모델로 사용했다.

3.2 3D-CvT

CvT에서는 하나의 프레임 정보만으로 컨볼루션 투영을 통해 키, 쿼리, 밸류를 생성했으나, 3D-CvT에서는 비디오 입력을 처리하기 위해 기존 2D-CNN을 입력 영상에 따라 3D 컨볼루션을 적용한 3D-CNN으로 변경한다. 3D-CvT의 전체 구조는 Fig. 3과 같다. 3D-CvT는 L개의 단계(Stage)로 구성되고, 각 단계는 컨볼루션 임베딩(Conv. Embed) 모듈과 N개의 층(Layer)으로 구성된다. 각 층은 컨볼루션 투영(Conv. Proj), 멀티 헤드 어텐션(MHSA)으로 구성된다. 본 논문에서는 3개의 단계와 각 단계별로 1, 2, 6개의 층이 있는 3D-CvT-(1,2,6) 모델과 4개의 단계와 각 단계별로 1, 2, 3, 2개의 층이 있는 3D-CvT-(1,2,3,2) 모델을 제안한다. 두 제안 모델의 자세한 구조는 Table 1에 나타내었다.

각 단계의 컨볼루션 임베딩 모듈에서는 입력 동영상으로부터 시간적으로 변하는 정보를 취득하기 위해 시간축으로 확장한 3D 커널을 사용하는데, 커널의 시간 축의 크기는 1단계에서는 7, 나머지 단계에서는 3을 사용한다. 각 단계의 MHSA 연산을 위해 3D-CvT는 컨볼루션 투영을 통해 쿼리, 키 밸류 값을 Fig. 5와 같이 계산한다. 컨볼루션 투영은 입력 X ($\in \mathbb{R}^{T\times C\times H\times W}$)와 커널 , 스트라이드((1,1,1), (2,2,1)) 변수에 대한 다음과 같은 함수로 정의할 수 있다.

$$Q = \text{3D-CNN}(X, \mathbb{K}_q, 1, 1, 1) \tag{1}$$

$$K, V = \text{3D-CNN}(X, \mathbb{K}_{k,v}, 2, 2, 1) \tag{2}$$

따라서, 쿼리의 크기는 T×C×H×W, 키와 밸류의 크기는 T×C×(H/2)×(W/2)로 결정된다. 계산된 키, 쿼리, 밸류값은 트랜스포머에서 일반적으로 사용되는 어텐션 연산을 통해 특징맵으로 변환되고, 다음 과정으로 전달된다.

3D-CvT 구조를 최적화하기 위해 각 설정값들은 Table 1과 같이 주의깊게 결정되었다. 1단계에서는 시간축 스트라이드를 2로 설정하여 입력 영상의 길이 T를 T/2로 줄인다. 1단계와 2단계 사이에는 3D 풀링(Max-Pooling)을 사용하여 T/2를 T/4로 줄인다. 이와 같이 시간축으로 특징 벡터의 크기를 줄임으로써 좀더 긴 수화 영상을 입력으로 사용할 수 있는 이득을 기대할 수 있다. 특징 맵의 X축 및 Y축으로의 크기를 줄이기 위해, 1단계 컨볼루션 임베딩에서는 X, Y의 스트라이드를 4, 1 단계 이후 컨볼루션 임베딩에서는 스트라이드를 2로 설정하였다.

Table 1. The Detailed Structure of 3D-CvT. Input: T*3*224*224 (T: The Number of Frames), Conv. Embed.: Convolution Embedding, Conv. Proj.: Convolution Projection, MHSA : Multihead Self Attention, H : The Number of Heads, D : Feature Dimension

	Output Size (T×D×H×W)	Modules	3D-CvT (1,2,6) & 3D-CvT (1,2,3,2) (Proposed)	
Stage-1	T/2 × 64 × 56 × 56	Conv. Embed	Kernel : 7 × 7 × 7, #Channel : 64, Stride : 4 × 4 × 2	
	T/2 × 64 × 56 × 56	Conv. Proj MHSA #Layers	Kernel : 3 × 3 × 3 H = 1, D = 64 ×1	Kernel : 3 × 3 × 3 H = 1, D = 64 ×1
Pooling	T/4 × 64 × 56 × 56	3D Max Pooling, Stride : 1 × 1 × 2		
Stage-2	T/4 × 192 × 28 × 28	Conv. Embed	Kernel : 3 × 3 × 3, #Channel : 192, Stride 2 × 2 × 1	
	T/4 × 192 × 28 × 28	Conv. Proj MHSA #Layers	Kernel : 3 × 3 × 3 H = 3, D = 192 ×2	Kernel : 3 × 3 × 3 H = 3, D = 192 ×2
Stage-3	T/4 × 384 × 14 × 14	Conv. Embed	Kernel : 3 × 3 × 3, #Channel : 384, Stride 2 × 2 × 1	
	T/4 × 384 × 14 × 14	Conv. Proj MHSA #Layers	Kernel : 3 × 3 × 3 H = 6, D = 384 ×6	Kernel : 3 × 3 × 3 H = 6, D = 384 ×3
Stage-4	T/4 × 768 × 14 × 14	Conv. Embed	-	Kernel : 3 × 3 × 3, #Channel : 768, Stride : 1 × 1 × 1
	T/4 × 768 × 14 × 14	Conv. Proj MHSA #Layers		Kernel : 3 × 3 × 3 H = 12, D = 768 ×2
Head	T/4 × 768 × 1	Mean		
	T/4 × 1066	Linear	1066	
GFLOPs			94.8	90.5

Table 2. Performance Comparison on PHOENIX2014T. The Performance of 3D-CvT is Comparable to the SoTA Models, Two-Stream (RGB and Landmark Inputs) and MMTL (RGB Inputs). The Best Performance is Marked in Bold and the Second Best Performance is Marked Underscored for Each Column.

	Model	Stage	Modality	Dev BLEU-1 ~ BLEU-4				Test BLEU-1 ~ BLEU-4			
2018	NSLT [11]	×	RGB	42.88	30.30	23.02	18.48	43.29	30.39	22.82	18.13
2020	SLT [2]			47.29	34.40	27.05	22.38	46.61	33.73	26.19	21.32
2020	STMC [3]			48.27	35.20	27.47	22.47	48.73	36.53	29.03	22.40
2022	MMTL [5]			53.95	41.12	33.14	27.61	53.97	41.75	33.84	28.39
2022	Two Stream [6]		RGB, Skeleton	**54.32**	**41.99**	**34.15**	**28.66**	**54.90**	**42.43**	**34.46**	**28.95**
3D-CvT (ours)		1,2,6	RGB	53.35	40.56	32.61	27.02	53.55	40.78	32.78	27.13
		1,2,3,2		53.49	40.72	32.94	27.23	53.72	40.92	33.08	27.35

Table 3. FLOPS Comparison with the SOTA Models. 3D-CvT Shows Lower FLOPs than MMTL.

Model	VFE	Stage	Modality	VFE FLOPs (100 frames)
MMTL[5]	S3D	-	RGB	250G
Two Stream [6]	S3D		RGB, Skeleton	-
3D-CvT (ours)		1,2,6	RGB	94.8G
		1,2,3,2		90.5G

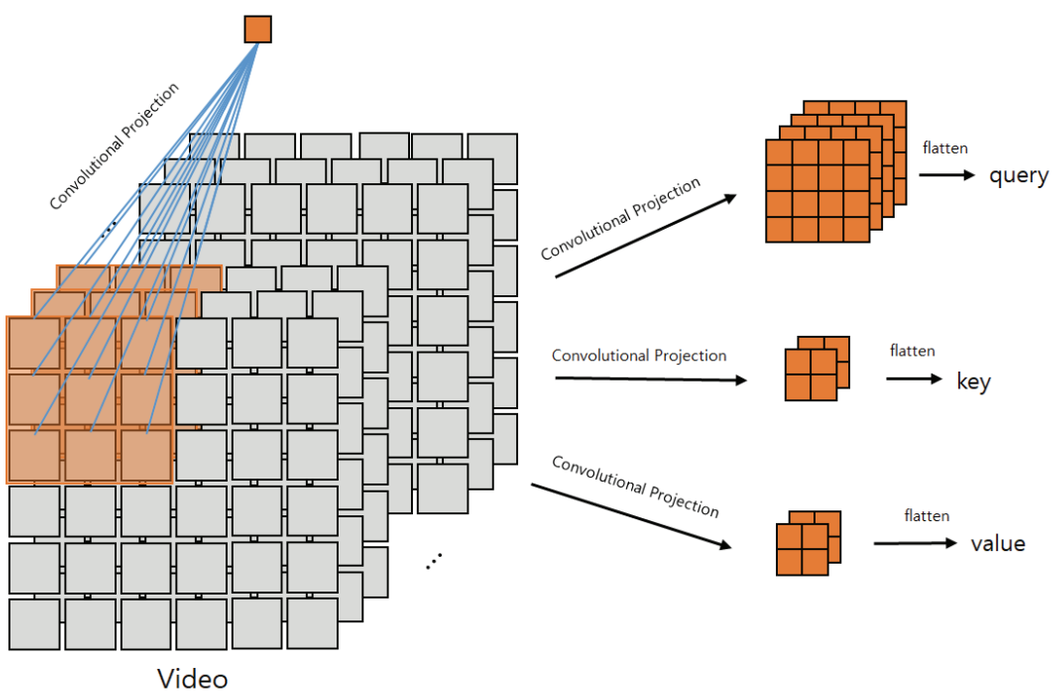


Fig. 5. The Convolution Projectio of 3D-CvT

3D-CvT-(1,2,6) 모델과 3D-CvT-(1,2,3,2) 모델은 3, 4 단계에서 차이가 있다. 3D-CvT-(1,2,6)는 3단계에서 6개의 층을 갖고, 3D-CvT-(1,2,3,2)는 3개의 층만을 가진다. 3단계만을 갖는 3D-CvT-(1,2,6)와는 달리 3D-CvT-(1,2,3,2)는 4번째 단계를 두어 네트워크를 더 깊게 설계했다. 3D-CvT-(1,2,6)에 비하여 3D-CvT(1,2,3,2)에 한 단계를 더 추가하여 채널의 크기를 늘림으로써 파라미터의 수도 두 배로 늘어났지만, 3단계와 4단계의 층 수를 줄임으로써, 연산량도 줄이는 효율성을 추구했다. 3D-CvT는 모든 단계(Stage) 후 마지막으로 선형 연산(Linear Layer)을 통해 T/4 × 1066 크기의 특징맵을 완성한다.

기존 CvT는 이미지 분류를 위해 설계된 모델로, 마지막 단

계에서 분류(cls) 토큰을 주로 사용한다. 반면, 3D-CvT는 분류가 아닌 수어소 배열을 출력하는 것이 목적이므로 분류 토큰 대신 최종 선형 연산 전에 특징 벡터의 평균을 통해 영상 프레임의 특징을 추출한다.

설계된 3D-CvT는 사전 학습을 거친 후, 본 학습을 진행하도록 했다. 먼저 동작 인식 데이터셋 Kinetics400 [11]으로 3D-CvT를 사전 학습하고, 수어 데이터셋 WLASL [12]를 사용하여 재학습하였다. 이렇게 사전 학습된 3D-CvT는 PHOENIX2014T [1] 데이터셋을 사용하여 본 학습을 진행하였다.

4. 실 험

4.1 데이터셋과 평가

1) 데이터셋

PHOENIX2014T [1]는 수화 번역 모델 [1,2,3,4,5,6]을 위해 널리 사용되는 데이터셋이다. 이 데이터셋은 독일 공영 텔레비전 방송국의 일일 뉴스와 일기 예보의 수어 통역 동영상으로 구성되어 있다. 7096개의 훈련 데이터와 519개, 642개의 평가 및 테스트 데이터가 포함되어 있으며, 1066개의 수어소와 2887개의 독일어 단어가 포함되어 있다. 각 데이셋은 비디오-수어소-구어의 쌍으로 구성되어 있다.

2) 평가 지표

제안 기술의 수어 번역 성능을 평가하기 위해 기존 Sign2Text 연구들[2,5]과 마찬가지로 BLEU 점수 [14]를 평가 지표로 사용했다. BLEU 점수는 기계 번역의 품질을 평가하는 지표로, 정답 문장과 예측 문장 사이의 n-그램 포함 정도를 나타낸다.

4.2 학습 방법

본 논문에서 모델의 학습은 최근 연구인 MMTL [5]의 사전 학습 방법을 따랐다.

1) VFE의 사전 학습

VFE의 사전 학습을 위해 3D-CvT를 행동분류 데이터셋인 Kinetics400을 사용하여 학습하고, 수어 단어 영상 데이터셋인 WLASL [12]을 사용하여 재학습했다. 마지막으로 수화 문장 데이터셋인 PHOENIX2014T [1]을 사용하여 Sign2Gloss 모델로 활용할 수 있도록 학습했다.

Sign2Gloss 모델로의 학습시에는 음성 인식 모델에서 흔히 사용되는 CTC 손실함수를 사용했다. VFE를 통해 얻은 수어소의 특징 벡터는 분류기와 소프트맥스를 통과하여 $p(\pi|V)$로 변환되는데, 여기서 π는 특정 수어소 배열 G를 얻기 위한 하나의 CTC 경로(Path)이고, V는 입력 비디오를 의미한다. 이때, 입력 V가 주어졌을 때, 수어소 배열 G를 얻는 확률 $p(G|V)$는 가능한 모든 CTC 경로 π를 매개변수로 하여 다음과 같이 계산한다:

$$p(G|V) = \sum_{\pi \in B} p(\pi|V) \tag{3}$$

여기서 B는 G를 얻기 위한 모든 가능한 CTC 경로의 집합이다. 이를 바탕으로 VFE의 손실함수는 다음과 같이 정의한다:

$$L_r = -\ln p(G \mid V) \tag{4}$$

2) 번역 모델 사전 학습

번역 모델의 구현은 MMTL[5] 방법을 따랐다. 번역 모델의 학습을 위해 25개 언어 말뭉치 데이터셋인 CC25를 사용하여 사전 학습된 mBART-large-cc25를 사용했다. 사전 학습된 mBART는 25개 청각 언어로 학습되었기 때문에, 이를 수어 번역에 활용하기 위하여 PHOENIX2014T 데이터로부터 독일어 단어의 임베딩을 생성하여 mBART를 재학습했다. 즉, 사전 학습된 mBART의 임베딩을 PHOENIX2014T을 따라 변경하여 Gloss2Text 모델로 역할할 수 있도록 학습했다. 학습에 필요한 손실함수는 MMTL과 마찬가지로 크로스 엔트로피 오차를 최소화하도록 설계했는데, 이에 해당하는 mBART의 손실함수 Lt는 다음 식(5)와 같다. 식에서 S는 번역 모델의 출력 텍스트 배열을 의미한다.

$$L_t = -\log p(S \mid G) \tag{5}$$

3) 종단 간 학습

본 논문에서는 Sign2Text를 학습시키기 전에 VFE와 번역 모델을 각각 별도로 사전 학습시켰다. 또, MMTL의 VL-mapper 구조와 같이 FC-MLP (Fully-Connected Multi-Layer Perceptron)를 사용하여 사전 학습된 두 모델을 연결했다. 이는 VFE의 출력인 수어소 특징을 번역 모델의 입력으로 사용하기에 적합한 특징으로 변환하기 위함이다. 종단 간 학습 중에는 VFE의 파라미터를 고정시키고 MLP와 번역 모델의 파라미터는 고정시키지 않았다.

전체 모델의 종단간 학습을 위해 두 종류의 손실함수를 사용한다. 첫째, 수어 비디오 입력을 받아 VFE를 통해 수어소 특징 벡터를 얻고, 이를 통해 VFE 손실함수인 Lr을 계산한다. 둘째, 수어소의 특징을 MLP와 번역 모델에 입력으로 전달하여 번역 손실함수인 Lt를 얻는다. 따라서, 전체 모델의 학습에 사용되는 손실함수는 다음과 같이 정의할 수 있다:

$$\mathcal{L} = \mathcal{L}r + \mathcal{L}t \tag{6}$$

4.3 성능 비교

Table 2, 3에 PHOENIX2014T에 대한 제안 모델의 번역 성능 평가를 요약했다. Table 2는 수어 번역에 있어 Two-Stream 모델과 MMTL 모델이 대체로 가장 좋은 성능을 보임을 나타낸다. MMTL는 입력으로 RGB 프레임만을 사용하고,

Two-Stream 모델은 입력으로 RGB 프레임과 수어 화자의 관절 위치(Landmark)를 표시한 히트맵을 동시에 활용한다. MMTL과 Two-Stream 모델은 CNN 기반 모델이다.

실험 결과 3D-CvT는 두 SoTA 모델에 비해 1 BLUE 점수 내외의 차이를 보인다. 즉, BLUE 점수로 나타내어지는 번역 성능에서 있어서 3D-CvT는 현재 SoTA 모델과 거의 유사한 성능을 보임을 알 수 있다.

Table 3은 3D-CvT와 타모델의 연산량을 나타낸다. 연산량이란 하나의 비디오 프레임이 입력되었을 때, 특징 벡터를 추출할때까지의 VFE의 정방향 연산양을 의미한다. RGB 비디오를 사용하는 MMLT는 250G FLOPs의 연산량을 필요로하는 반면, 3D-CvT는 90G FLOPs 내외의 적은 연산량을 필요로 한다. 3D-CvT의 적은 연산량은 3D 컨볼루션 연산이 적은 파라미터를 사용하면서도 수어 번역에 필요한 시공간 특징을 효과적으로 추출할 수 있는데 기인한다.

본 실험을 통해 제안한 3D-CvT는 수어 영상을 입력으로 받는, 즉 같은 형태(Modality)의 입력을 사용하는 최근 모델 MMTL과 비교했을 때, 더 적은 연산량(FLOPs)으로도 타 SoTA 모델과 비슷한 성능을 보임을 증명하였다. 각 모델의 번역 성능과 연산량은 Fig. 1에 함께 나타낸바 있다.

5. 결 론

본 논문에서는 수어 번역 모델의 VFE 부분을 개선하기 위해 이미지 인식 모델인 CvT를 동영상으로 확장한 3D-CvT 모델을 제안했다. 3D-CvT 모델은 CNN의 컨볼루션 연산과 트랜스포머의 어텐션 연산의 장점을 결합한 모델로 수어 영상에 대한 기계 번역을 위해 효과적인 시공간 특징 정보의 추출에 적합하다. 또한 3D-CvT는 더 적은 연산량으로도 기존 모델과 유사한 번역 성능을 보임을 실험적으로 증명하였다. 향후에는 3D-CvT의 성능을 높이기 위한 3D-CvT의 다양한 변형에 대해 고찰할 계획이다.

References

[1] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, “Neural sign language translation,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[2] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign language transformers: Joint end-to-end sign language recognition and translation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[3] K. Yin, and R. Jesse, “Better sign language translation with STMC-transformer,” *arXiv preprint arXiv:2004.00588*, 2020.

[4] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, “Improving sign language translation with monolingual data by sign back-translation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[5] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, “A simple multi-modality transfer learning baseline for sign language translation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[6] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, “Two-stream network for sign language recognition and translation,” *Advances in Neural Information Processing Systems*, Vol.35, pp.17043-17056, 2022.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[8] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan et al., “Cvt: Introducing convolutions to vision transformers,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[9] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[10] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, “Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification,” *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[11] W. Kay et al., “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.

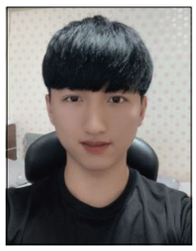
[12] D. Li, C. R. Opazo, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

[13] Y. Liu et al., “Multilingual denoising pre-training for neural machine translation,” *Transactions of the Association for Computational Linguistics*, Vol.8, pp.726-742, 2020.

[14] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[15] Y. Wang et al., “Internvideo: General video foundation models via generative and discriminative learning,” *arXiv preprint arXiv:2212.03191*, 2022.

[16] A. J. Piergiovanni, W. Kuo, and A. Angelova, "Rethinking video vits: Sparse video tubes for joint image and video learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
[17] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," *ICML*, Vol.2, No.3, 2021.
[18] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "Vivit: A video vision transformer," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
[19] M. Lewis et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.
[20] J. Guo et al., "Cmt: Convolutional neural networks meet vision transformers," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
[21] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009.



성 호 렬

https://orcid.org/0000-0002-8132-4183
e-mail : supernova817@korea.ac.kr
2021년 고려대학교 컴퓨터정보학과(학사)
2023년 고려대학교 컴퓨터정보학과(석사)
관심분야 : 수어 인식, 수어 번역, 액션 인식

조 현 중

https://orcid.org/0000-0003-1487-895X
e-mail : raycho@korea.ac.kr
1996년 경북대학교 전자공학부(학사)
1998년 포항공과대학교 전자전기공학(석사)
2006년 미국 버지니아 공과대학교 컴퓨터공학(박사)
2009년 ~ 현 재 고려대학교 컴퓨터융합소프트웨어학과 교수
관심분야 : Machine Learning Techniques, Action/Gesture Recognition, Sign Language Translation, Application Using Optimization Theory, Cyber-physical Systems