비디오 캡셔닝을 적용한 수어 번역 및 행동 인식을 적용한 수어 인식

김기덕O, 이근후*
*(주)쓰리아이퓨처,
O(주)쓰리아이퓨처
e-mail: lghoo@naver.com*, kimsjpk@naver.comO

Sign language translation using video captioning and sign language recognition using action recognition

Gi-Duk KimO, Geun-Hoo Lee*
*3IFuture,
O3IFuture

● 요 약 ●

본 논문에서는 비디오 캡셔닝 알고리즘을 적용한 수어 번역 및 행동 인식 알고리즘을 적용한 수어 인식 알고리즘을 제안한다. 본 논문에 사용된 비디오 캡셔닝 알고리즘으로 40개의 연속된 입력 데이터 프레임을 CNN 네트워크를 통해 임베딩 하고 트랜스포머의 입력으로 하여 문장을 출력하였다. 행동 인식 알고리즘은 랜덤 샘플링을 하여 한 영상에 40개의 인덱스에서 40개의 연속된 데이터에 CNN 네트워크를 통해 임베딩 하고 GRU, 트랜스포머를 결합한 RNN 모델을 통해 인식 결과를 출력하였다. 수어 번역에서 BLEU-4의 경우 7.85, CIDEr는 53.12를 얻었고 수어 인식으로 96.26%의 인식 정확도를 얻었다.

키워드: 수어 번역(sign language translation), 수어 인식(sign language recognition), 비디오 캡셔닝(video captioning), 행동 인식(action recognition)

I. Introduction

청각 장애인의 경우 수어를 사용하여 타인과 의사소통하는데 이때 수어를 알지 못하면 의사소통에 제약받게 된다. 장애인에게도 평등한 사회를 제공하기 위해 수어 번역 및 수어 인식 시스템의 필요성이 증대되고 있다. 수어 번역의 경우 전체 길이에서 일정 길이를 단어로 인식하고 이를 모아서 번역하는 방법을 주로 사용한다. 이 경우 수어 번역 데이터세트에 대한 라벨링의 비용이 증가하게 된다. 본 논문에서는 기존의 비디오 캡셔닝 알고리즘을 사용하여 전체 문장을 학습시킴으로써 데이터 라벨링의 비용을 낮춘 수어 번역 알고리즘을 제안한다. 비록 성능은 낮게 나왔으나 이는 입력데이터를 가공하지 않고 CNN을 거친 임베딩 데이터에 트랜스포머를 적용한 가장 간단한 형태의 비디오 캡셔닝을 적용하여 효율적으로 수어 번역의 결과를 얻었다. 수어 인식의 경우 행동 인식에 사용된 CNN-RNN 구조를 사용하고 랜덤 샘플링을 통한 데이터 증대 방법을 적용하여 간단한 딥러닝 네트워크를 사용하여 높은 정확도를 얻을 수 있었다. Kim 등[1]은 수어 번역의 정확도를 높이고자 입력 데이터가 많을 때는 랜덤 샘플링을 하고 적을 때는 확률 분포를 기반으로 프레임을 끼워넣는 데이터 증대를 사용하여 BLEU 점수를 높였다. 다양한 데이터 전처리 방법과 기존에 연구된 비디오 캡셔닝 방법을 적용한다면 더욱 정확한 수어 번역 결과를 얻을 수 있을 것이다. 2장에서는 비디오 캡셔닝을 적용한 수어 번역 연구 흐름과 본 논문에서 사용된 데이터세트에 대한 설명을 하고 3장에서는 실험 결과에 대해서 서술한다. 4장에서는 결론을 서술한다.

II. Preliminaries

1. Related works

1.1 비디오 캡셔닝을 적용한 수어 번역

데이터 전처리 방법으로 입력 이미지에 포즈 추정[2]을 적용하여 인물의 특징을 추출한 이미지를 입력으로 수어 번역을 적용하였다. 네트워크 구조로는 CNN-RNN의 인코더 디코더 방법[3]을 결합한 구조를 적용하는 방법과 RNN 대신에 Bert 기반의 수어 번역 알고리즘[4] 방법이 사용되었다. 그리고 GCN[5](Graph Convolution Network)를 적용한 방법이 연구되었다.

1.2 행동 인식을 적용한 수어 인식

스포츠 분석, 이상행동 탐지를 위해 행동 인식이 연구되었다. Simonyan 등[6]은 본 논문의 방법과 유사하게 랜덤 샘플링을 하고 일정 길이만큼 슬라이싱 한 공간 데이터와 시간 데이터에 3D ConvNet을 적용하였다. LRCN(Long term Recurrent Convolution Network)[7]에서는 CNN을 통한 이미지 임베딩 데이터에 RNN을 통하여 행동 인식 클래스를 출력하는 네트워크를 제안하였다.

2.1 수어 번역 데이터세트

본 논문에서는 독일에서 만든 RWTH-PHOENIX-Weather 2014-T[8] 데이터세트를 사용하였다. 이 외에도 미국에서 만든 ASLG-PC12[9] 데이터세트, 한국에서 만든 Korean sign language dataset KETI[10]가 있다. RWTH-PHOENIX-Weather 2014-T 데이터세트의 경우 2009년부터 2011년의 3년 동안 기상 뉴스의 수어 화면을 초당 25 프레임의 210 x 260 크기로 저장하였다. 영상에 대한 독일어 자막이 달려 있으며 학습 데이터의 경우 7,095개 영상, 테스트 데이터의 경우 641개 영상과 독일어 자막으로 구성되어 있다.

2.2 수어 인식 데이터세트

본 논문에서는 KSL 데이터세트[11]를 사용하였다. 20명의 사람이 77개의 단어에 대한 수어 영상으로 구성되어 있으며 영상에는 17곳의 일상생활 배경을 포함한다. 1,229개의 영상으로 총 112,564 개의 1280 x 720 프레임이 저장되었다. RGB 영상과 optical flow 영상이 있으나 본 논문에서는 RGB 영상만을 사용하였다.

III. The Proposed Scheme

본 논문에서는 수어 번역의 경우 입력 영상을 ImageNet에 사전 학습된 VIT(Vision Transformer)[12]를 사용하여 이미지를 768개의 벡터로 임베딩 하였다. 임베딩 된 벡터를 입력으로 트랜스포머에 넣어서 수어 번역 데이터세트의 입력을 비교하여 학습하였다. RWTH-PHOENIX-Weather 2014-T 데이터세트에서 독일어로 된 캡션은 구글 번역기를 사용하여 영어로 번역하고 학습을 진행하였다. 학습 네트워크로 디코더를 MLP(Multi Layer Perceptron)과 트랜스포머를 사용하여 성능을 비교하였다. 성능은 표 1과 같다.

Table 1. 수어 번역 데이터세트 알고리즘 성능 비교

알고리즘	Bleu 4	ROUGE-L
Kim 등[1]	**13.31**	24.72
VIT + MLP	7.15	**22.26**
VIT + transformer	7.85	25.40
알고리즘	METEOR	CIDEr
Kim 등[1]	**25.86**	-
VIT + MLP	9.14	36.93
VIT + transformer	12.44	53.12

수어 인식의 경우 영상에서 임의의 인덱스 40개에 대해 길이 40만큼 슬라이싱하고 VIT를 통해 임베딩 한 데이터에 GRU와 트랜스포머 모델을 합친 간단한 구조의 RNN 모델에 학습하였다. 모델의 구조는 그림 1과 같다. 그리고 성능은 표 2와 같다.

```
model = Sequential()
model.add(Conv1D(filters=384, kernel_size=3,
          padding='same', activation='relu'))
model.add(Bidirectional(GRU(384, return_sequences=True)))
transformer_block = TransformerBlock(768, 4, 384)
model.add(transformer_block)
model.add(layers.GlobalAveragePooling1D())
model.add(layers.Dropout(0.1))
model.add(Dense(NUM_CLASSES, activation='softmax'))
```

Fig. 1. 본 논문에 사용된 수어 인식 모델

Table 2. 수어 인식 데이터세트 알고리즘 성능 비교

알고리즘	RGB	Flow
TSN[14]	-	79.80
LRCN	27.06	54.14
VIT + RNN	**96.26**	-

IV. Conclusions

본 논문에서는 기존의 연구된 비디오 캡셔닝과 행동 인식을 사용하여 수어 번역과 수어 인식에 대한 알고리즘을 제안한다. 알고리즘의 경우 이미지에 대한 전처리 없이 임베딩 후 디코딩하는 간단한 구조를 적용하였다. Kim [1] 등은 확률론을 기반으로 이미지 증대를 하여 BLEU-4 값은 증가하였지만, ROUGH-L 값과 METEOR 값은 조합에 따라 감소하는 경우가 있었다. 앞으로 다른 논문에서 제안된 데이터 전처리 방법을 사용하여 수어 번역과 수어 인식의 성능을 높일 예정이다. 논문에 사용된 코드는 블로그에 올려두었다. (https://blog.naver.com/kimsjpk/223283505844)

REFERENCES

[1] KIM, Youngmin, et al. Keypoint based sign language translation without glosses. arXiv preprint arXiv:2204.10511, 2022.

[2] CAO, Zhe, et al. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 7291-7299.

[3] BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[4] LU, Kevin, et al. Frozen pretrained transformers as universal computation engines. In: Proceedings of the AAAI

Conference on Artificial Intelligence. 2022. p. 7628-7636.
[5] KIPF, Thomas N.; WELLING, Max. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.
[6] SIMONYAN, Karen; ZISSERMAN, Andrew. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 2014, 27.
[7] DONAHUE, Jeffrey, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 2625-2634.
[8] CAMGOZ, Necati Cihan, et al. Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7784-7793.
[9] OTHMAN, Achraf; JEMNI, Mohamed. English-asl gloss parallel corpus 2012: Aslg-pc12. In: sign-lang@ LREC 2012. European Language Resources Association (ELRA), 2012. p. 151-154.
[10] KO, Sang-Ki, et al. Neural sign language translation based on human keypoint estimation. Applied sciences, 2019, 9.13: 2683.
[11] YANG, Seunghan, et al. The Korean sign language dataset for action recognition. In: International conference on multimedia modeling. Cham: Springer International Publishing, 2019. p. 532-542.
[12] DOSOVITSKIY, Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.