2-1-18-MMP

SpeechCLIP: Integrating Speech with Pre-Trained Vision and Language Model

Yi-Jen Shih[1], Hsuan-Fu Wang[1], Heng-Jui Chang[1,2], Layne Berry[3], Hung-yi Lee[1], David Harwath[3]

[1]National Taiwan University [2]MIT CSAIL [3]The University of Texas at Austin

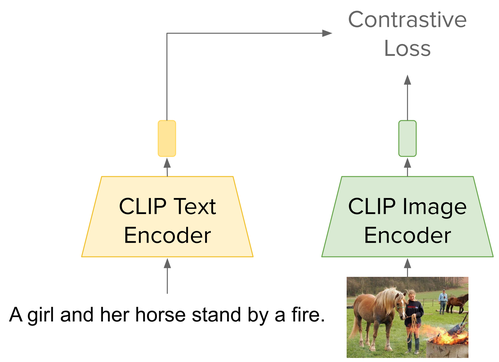
SLT 2022

Motivation

- Paired speech-image data is cheaper and easier to collect compared to transcription data
- Compared to previous works, SpeechCLIP utilize visually grounded speech and indirect text signals for training

Methods

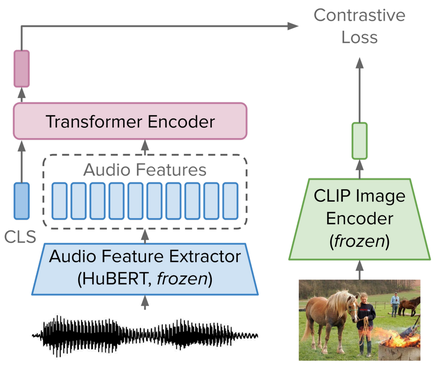
- CLIP [1]
  - Align image and speech into same embedding space and able to transfer to Computer Vision Tasks



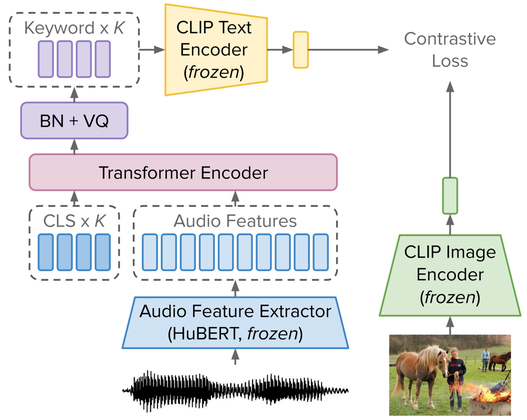
- SpeechCLIP
  - **Parallel**: the spoken utterance embedding is aligned with CLIP image encoder's output
  - **Cascaded**: *K* keywords are selected and inserted into CLIP text encoder to align with CLIP image encoder's output

Model	Audio Encoder	CLIP Image Encoder	Trainable Params (M)	Total Params (M)
Base	HuBBERT Base (95M)	ViT-B/32 (250M)	2.8 - 7.5	252 - 257
Large	HuBERT Large (316M)	ViT-L/14 (422M)	6.1 - 13.4	765 - 772

Model Details



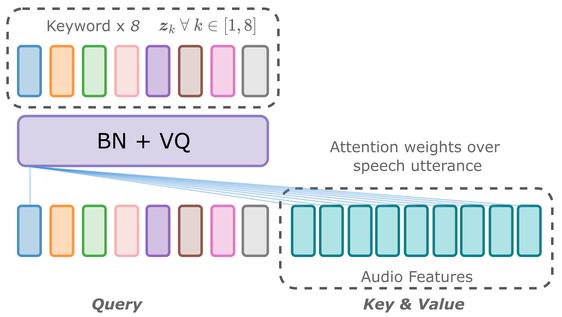
(a) Parallel SpeechCLIP

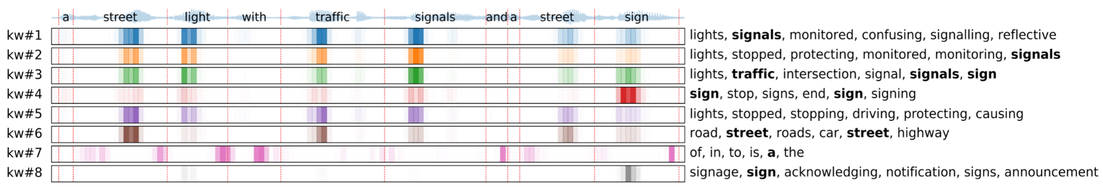


(b) Cascaded SpeechCLIP

Results

- Keyword Discovery
  - Keywords are retrieved by finding subwords with the highest cosine similarities between z_k and the corresponding subword embeddings





Model	kw1	kw2	kw3	kw4	kw5	kw6	kw7	kw8	Avg
Base (Flickr)	57.0	25.6	20.2	5.0	20.0	26.5	10.5	16.6	22.7
Large (Flickr)	56.5	19.6	20.5	37.5	21.7	34.6	26.4	44.7	32.7
Large (COCO)	27.5	22.4	35.8	61.0	21.6	54.2	60.1	22.9	38.2

Keyword hit rates for cascaded SpeechCLIP. Avg denotes averaged hit rate

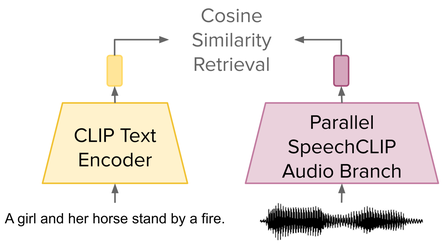
Reference

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision,", *ICML*, 2021.

[2] P. Peng and D. Harwath, "Fast-slow transformer for visually grounding speech,", *ICASSP*, 2022.

Results

- Zero-shot Speech-text Retrieval



	Speech → Text			Text → Speech		
Model	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.02	0.10	0.20	0.02	0.10	0.20
Parallel Large	60.32	81.81	88.18	65.45	85.82	91.27
Parallel Large (Sup.)	95.02	99.46	99.78	95.35	99.68	99.93

- Speech-Image Retrieval

	Speech → Image			Image → Speech		
Model	R@1	R@5	R@10	R@1	R@5	R@10
FaST-VGS$_{CO}$ [2]	31.8	62.5	75.0	42.5	73.7	84.9
FaST-VGS$_{CTF}$ [2]	**35.9**	66.3	77.9	48.8	78.2	87.0
Parallel Large	35.8	**66.5**	**78.0**	**50.6**	**80.9**	**89.1**
Cascaded Large	6.4	20.7	31.0	9.6	27.7	39.7

Conclusion

- SpeechCLIP established a new research direction of indirectly supervising speech models with text via other modalities.

Acknowledgement

- This work was supported by JSALT 2022
- We thank Taiwan Web Service(TWS) for providing computational resource

More Info

- Checkout our arXiv, code and blog for more info