中国科学院 CHINESE ACADEMY OF SCIENCES

中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY ,CHINESE ACADEMY OF SCIENCES

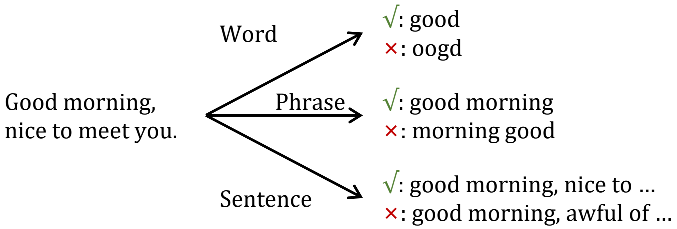
BMVC 2023

UniLip: Learning Visual-Textual Mapping with Uni-Modal Data for Lip Reading

Bingquan Xia, Shuang Yang, Shiguang Shan, Xilin Chen

Background

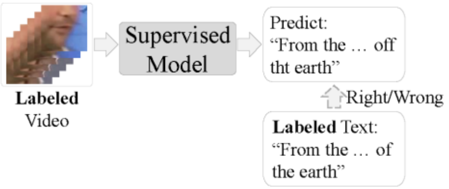
- Existing lip reading methods rely on large-scale labelled video-text pairs to perform supervised training.
- Collecting labeled video-text pairs are time-consuming, while collecting uni-modal videos and uni-modal texts are much easier.
- Uni-modal texts contain rich linguistic prior information of the target language which could facilitate lip reading.



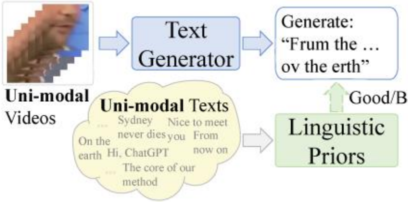
An example of linguistic priors

Motivation

- Utilize uni-modal videos and uni-modal texts to perform lip reading.



Supervised Approach



Our Approach

Video&Text Data Examples

Video

- LRS3: TED talks, 433h.
- LRS2: BBC shows, 224h.
- Vox2-433h: English sub-set of VoxCeleb2, 433h.

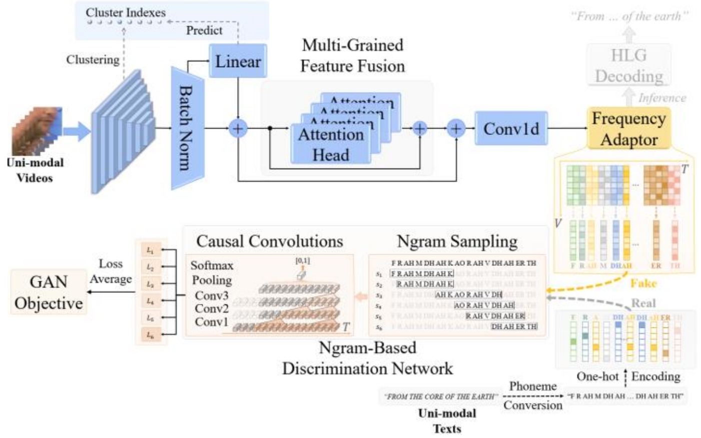
Text

- LRS3: 0.18M (M: Million utts).
- LRS2: 0.14M.
- TEDLIUM-v3: 0.27M, TED.
- Cantab-TEDLIUM: 7M, TED.
- LibriSpeech: 0.29M, audiobooks.

Texts are from rich sources and domains.

The proposed UniLip

- Decompose lip reading into two sub-tasks: (S1) learn linguistic priors from uni-modal texts (language modelling); (S2) generate text distributions conditioned on uni-modal videos (conditional generation).
- Propose a unified adversarial training framework to finish both (S1)and(S2).
- (S1): $\mathcal{D}$ maximizes the log likelihood of real samples; (S2): $\mathcal{G}$ generates text distributions that could deceive $\mathcal{D}$ conditioned on visual inputs.



- Multi-grained Learning of Linguistic Priors: alleviate the biases of text sources and domains by ngram sampling.
- Multi-grained Visual-Textual Mapping: adapt features of pre-trained models by integrating both local information and the global context.

Unsupervised Results

- UniLip's performance scales with the size of texts.
- UniLip can effectively accommodate videos and texts from different sources.

Training Video	Training Text	Test WER/% (↓) (Constrained)	Test WER/% (↓) (Unconstrained)
LRS3	LRS3	-	51.9(-)
	TEDLIUM	**51.2**	53.1(1.9↑)
	Cantab	61.8	60.8(1.0↓)
	LibriSpeech	N/A	64.9(∞↓)
LRS2	LRS2	-	**57.2**(-)
	LRS3	59.7	57.8(1.9↓)
	TEDLIUM	58.3	57.3(1.0↓)
	Cantab	60.7	58.9(1.8↓)
	LibriSpeech	N/A	N/A

Semi-supervised Results

- $L = L_{seq2seq} + \alpha L_{GAN}$.
- UniLip could effectively incorporate extra uni-modal data into the popular supervised Seq2Seq framework.

Labeled Hours/h	**Uni-modal Videos**	**Uni-modal Texts**	**Test WER/% (↓) (Base)**	**Test WER/% (↓) (Large)**
		LRS2		
224	——		30.6[39]	24.3[39]
			32.0*	28.1*
	LRS2	LRS2	31.2 (0.8↓)	27.8 (0.3↓)
	Vox2-433h	TEDLIUM	31.0 (1.0↓)	27.7 (0.4↓)
30	——		42.6[39]	31.6[39]
			42.0*	35.5*
	LRS2	TEDLIUM	41.1 (0.9↓)	34.0 (1.5↓)
433	——		32.4[39]	28.4[39]
			36.6*	32.6*
	LRS3	LRS3	35.4(1.2↓)	31.7 (0.9↓)
	Vox2-433h	TEDLIUM	N/A	31.5 (1.1↓)
	LRS2	TEDLIUM	36.2 (0.4↓)	N/A

*: our reproduced baselines

Visualization

- perform phoneme-level decoding and retrieve corresponding input lip images.
- UniLip successfully maps different phonemes to different lip shapes, such as "CH" and "M“.