Training Audio Captioning Models without Audio

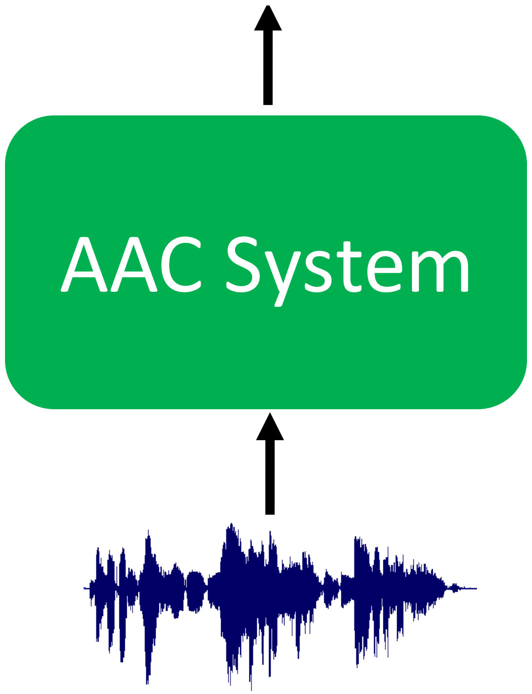
[1,2]Soham Deshmukh, [1]Benjamin Elizalde, [1]Dimitra Emmanouilidou,
[2]Bhiksha Raj, [2]Rita Singh, [1]Huaming Wang
[1]Microsoft, [2]Carnegie Mellon University

Overview

Leverage the multimodal space from a Contrastive Language Audio Pretraining (CLAP) to train an Audio Captioning system with **only text**.

Automated Audio Captioning (AAC) generates a description given an audio stream.

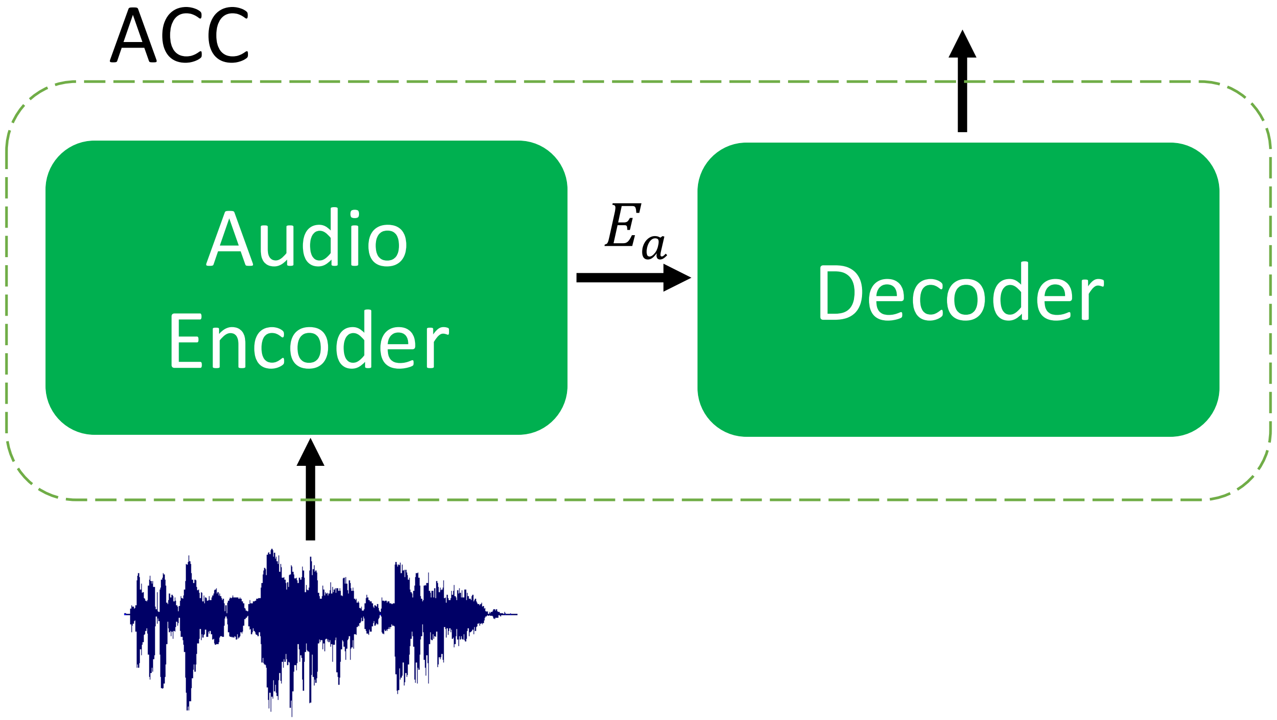
Children playing with birds chirping in the background



Different from Speech Transcriptions and Closed Captioning

Baseline architecture for AAC

Children playing with birds chirping in the background

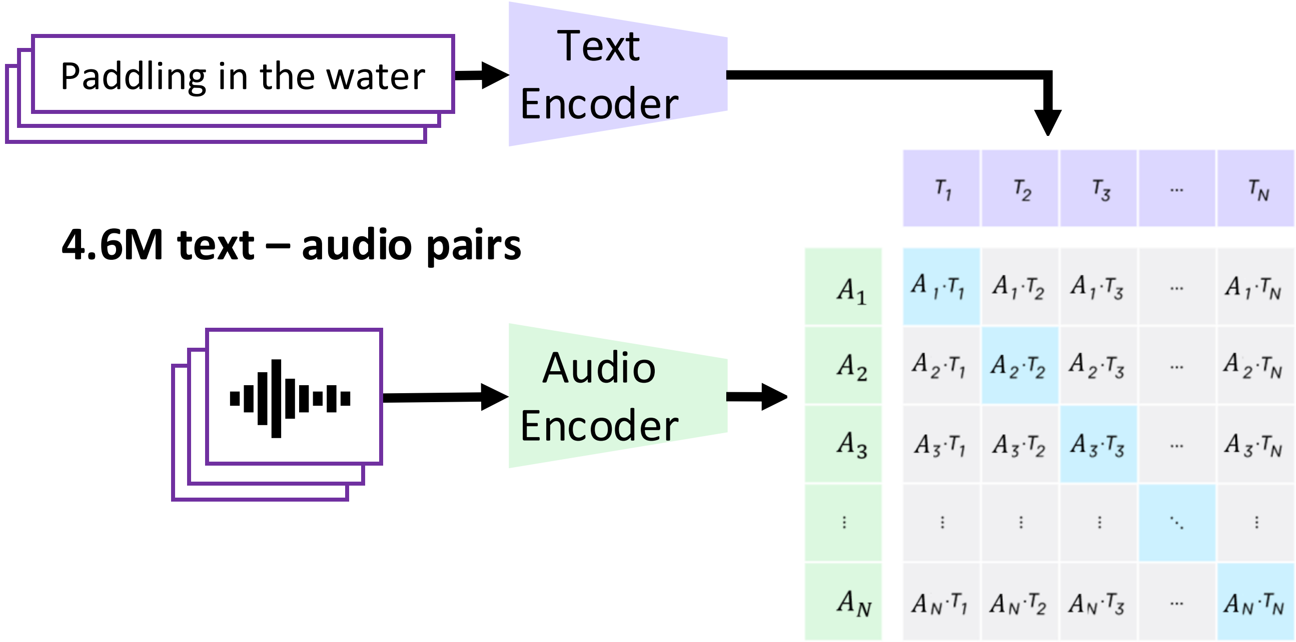


Training and Inference need an audio encoder

Contrastive Language-Audio Pretraining (CLAP)

Pretraining stage

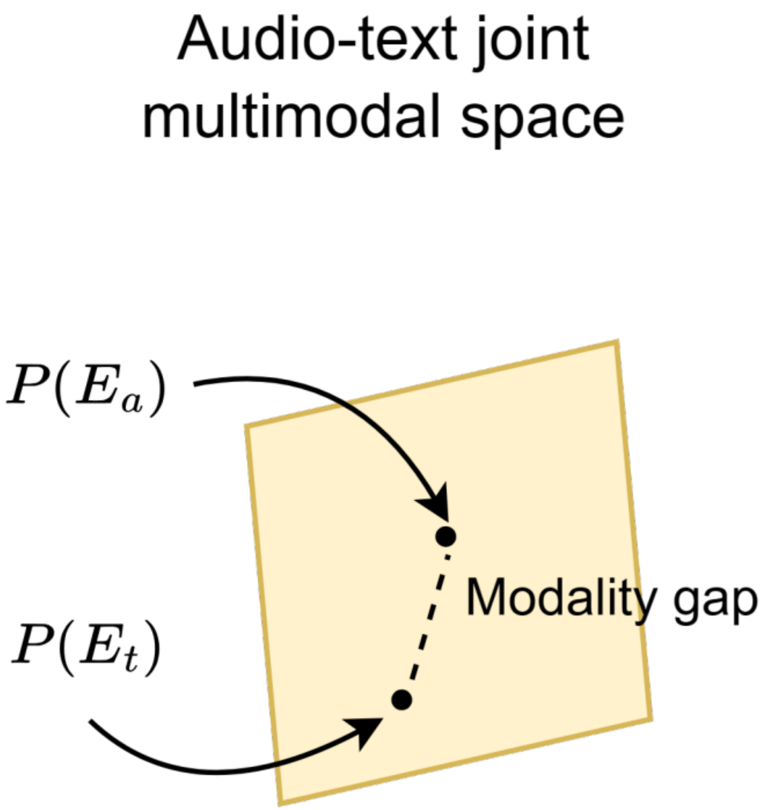
4.6M text – audio pairs



Minimize
average cross-entropy
between audio and text

Benjamin Elizalde, Soham Deshmukh, et al, "Natural language supervision for general-purpose audio representations," ICASSP 2024

Multimodal space of CLAP



An AAC model learns $P(Caption|E_a)$

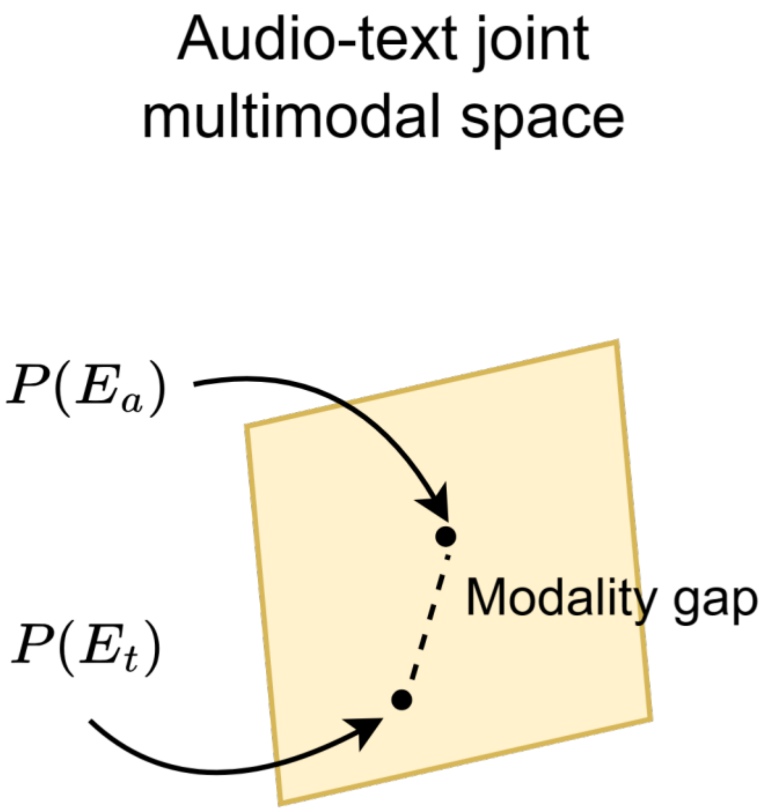
In CLAP, $P(Caption|E_a) = P(Caption|E_t)$

Therefore, for AAC, we can instead learn $P(Caption|E_t)$

Implications:

1. Use CLAP's text encoder for training and the audio encoder for inference.
2. Enable text-only training, no need for aligned audio and caption.

However, there exists a modality gap

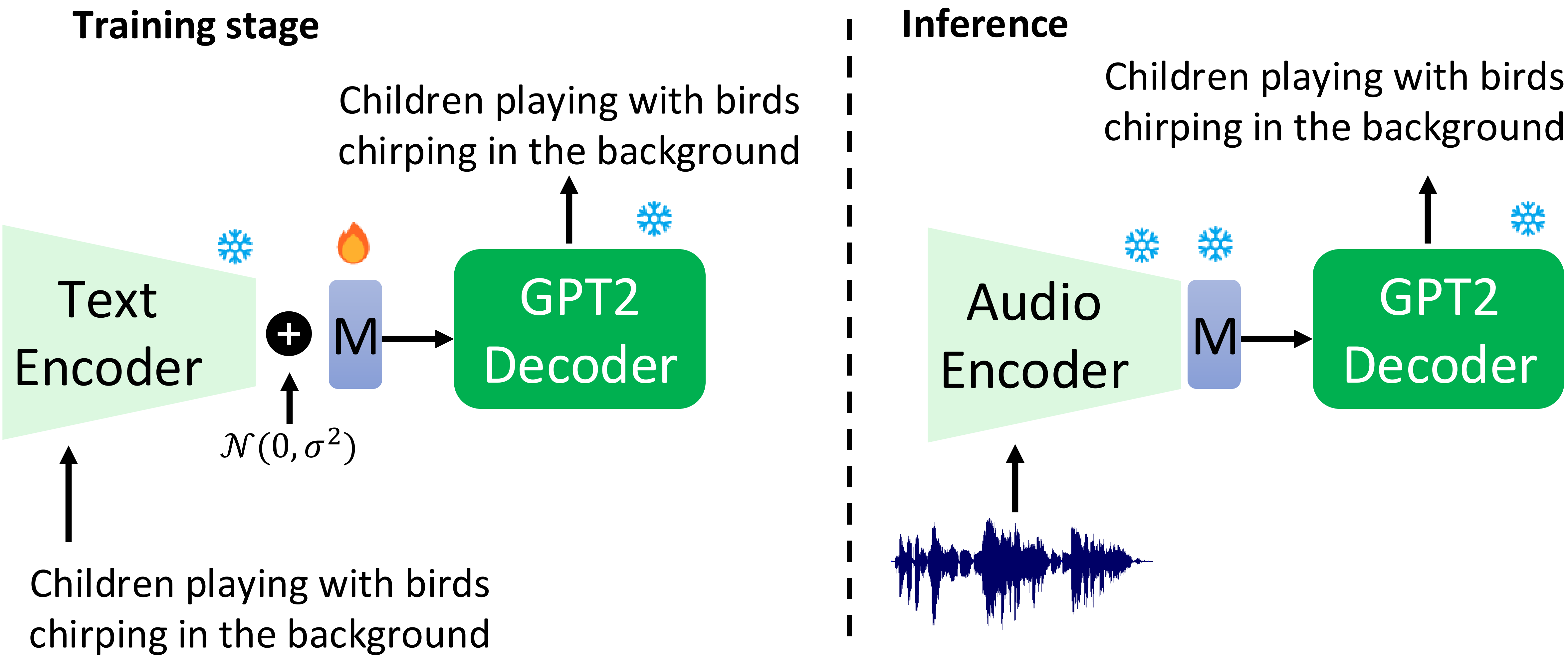


In practice,
$P(Caption|E_t) \neq P(Caption|E_a)$
instead,
$P(Caption|E_t) \sim P(Caption|E_a)$

The gap **limits** the direct swap of audio and text encoders.

To **bridge the gap**, we added zero-mean Gaussian noise at training

Proposed Text-only training for AAC



AAC with CLAP's audio encoder for audio-text training

Model	Eval. dataset	$BLUE_1$	$BLUE_2$	$BLUE_3$	$BLUE_4$	METEOR	$ROUGE_L$	CIDEr	SPICE	SPIDEr
Chen et al.	AudioCaps	0.489	0.292	0.178	0.106	0.152	0.346	0.265	0.093	0.179
Gontier et al.	AudioCaps	0.635	0.461	0.322	0.219	0.208	0.450	0.612	0.153	0.383
Mei et al.	AudioCaps	0.682	0.507	0.369	0.266	0.238	0.488	0.701	0.166	0.434
Kim et al.	AudioCaps	0.708	0.547	0.402	0.283	0.238	0.499	0.710	0.167	0.438
Audio-text (proposed)	AudioCaps	0.647	0.480	0.337	0.223	0.223	0.462	0.729	0.181	0.455
Chen et al.	Clotho	0.516	0.325	0.215	0.141	0.153	0.350	0.314	0.102	0.208
Gontier et al.	Clotho	0.461	0.282	0.182	0.117	0.136	0.318	0.251	0.083	0.167
Mei et al.	Clotho	0.516	0.318	0.204	0.127	0.157	0.351	0.313	0.105	0.209
Kim et al.	Clotho	0.539	0.346	0.227	0.142	0.159	0.366	0.319	0.111	0.215
Audio-text (proposed)	Clotho	0.574	0.375	0.250	0.155	0.173	0.381	0.398	0.123	0.261

We achieved SoTA performance in both datasets

AAC with CLAP's text encoder for text-only training

Model	Eval. dataset	$BLUE_1$	$BLUE_2$	$BLUE_3$	$BLUE_4$	METEOR	$ROUGE_L$	CIDEr	SPICE	SPIDEr
Chen et al.	AudioCaps	0.489	0.292	0.178	0.106	0.152	0.346	0.265	0.093	0.179
Gontier et al.	AudioCaps	0.635	0.461	0.322	0.219	0.208	0.450	0.612	0.153	0.383
Mei et al.	AudioCaps	0.682	0.507	0.369	0.266	0.238	0.488	0.701	0.166	0.434
Kim et al.	AudioCaps	0.708	0.547	0.402	0.283	0.238	0.499	0.710	0.167	0.438
Text-only (proposed)	AudioCaps	0.645	0.481	0.338	0.227	0.220	0.458	0.697	**0.178**	0.437
Audio-text (proposed)	AudioCaps	0.647	0.480	0.337	0.223	0.223	0.462	0.729	0.181	0.455
Chen et al.	Clotho	0.516	0.325	0.215	0.141	0.153	0.350	0.314	0.102	0.208
Gontier et al.	Clotho	0.461	0.282	0.182	0.117	0.136	0.318	0.251	0.083	0.167
Mei et al.	Clotho	0.516	0.318	0.204	0.127	0.157	0.351	0.313	0.105	0.209
Kim et al.	Clotho	0.539	0.346	0.227	0.142	0.159	0.366	0.319	0.111	0.215
Text-only (proposed)	Clotho	0.524	0.339	0.222	0.136	**0.173**	**0.371**	**0.379**	**0.132**	**0.256**
Audio-text (proposed)	Clotho	0.574	0.375	0.250	0.155	0.173	0.381	0.398	0.123	0.261

Achieves comparable performance to traditional audio-text training

Training with additional ~400k LLM generated captions (WavCaps)

Model	Eval. dataset	$BLUE_1$	$BLUE_2$	$BLUE_3$	$BLUE_4$	METEOR	$ROUGE_L$	CIDEr	SPICE	SPIDEr
Text-only	AudioCaps	0.645	0.481	0.338	0.227	0.220	0.458	0.696	0.178	0.437
Text-only†	AudioCaps	**0.653**	**0.484**	**0.342**	**0.232**	**0.226**	**0.459**	**0.697**	**0.179**	**0.438**
Text-only	Clotho	0.524	0.339	0.222	0.136	0.173	0.371	0.379	0.132	0.256
Text-only†	Clotho	**0.530**	**0.342**	**0.224**	**0.143**	0.164	0.367	0.377	0.117	0.247

Performance improved overall, especially on n-gram matching metrics

Xinhao Mei, et al, “WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research,” 2023

Infusing style with stylized captions

Train Dataset	Eval. dataset	BLUE_1	BLUE_2	SPIDEr
Original Clotho	Humor Clotho	0.370	0.162	0.092
Humor Clotho	Humor Clotho	**0.410**	**0.214**	**0.102**

Original: "Sand is being shoveled and dumped on the ground",
Humorous: "Sand relocation program: from shovel to ground, it's a gritty story".

Text-only training enables faster adaptation to different styles

Takeaways

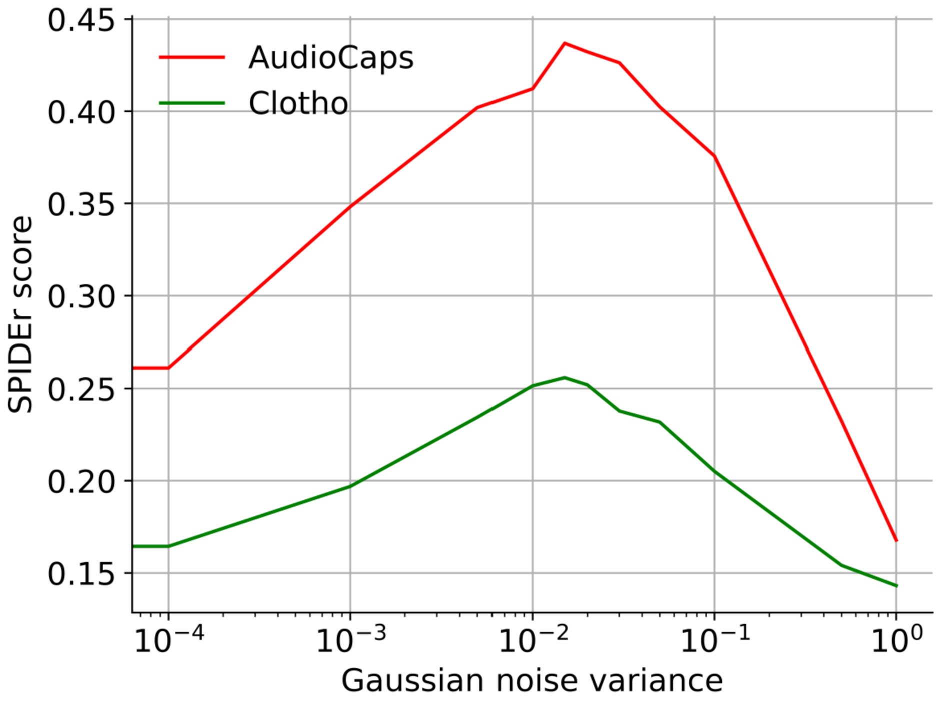
1. We introduced a text-only training approach for AAC.
2. We leveraged from the multimodal space learned by contrastive models.
3. Our text-only training achieves competitive results with the SoTA, while enabling improvement and stylization of captions.

Soham Deshmukh, Benjamin Elizalde

{sdeshmukh, benjaminm}@microsoft.com

Appendix

Effect of variance of Gaussian Noise



The variance of Gaussian noise can be approximated by:

infinity norm between the audio and text embeddings of randomly chosen examples

There exists a better value of noise variance irrespective of target dataset

Gaussian noise vs trained linear adapter

Model	Adapter	Eval. dataset	$BLUE_1$	$BLUE_2$	$BLUE_3$	$BLUE_4$	METEOR	$ROUGE_L$	CIDEr	SPICE	SPIDEr
Text-only	Gaussian	AudioCaps	**0.645**	**0.481**	**0.338**	**0.227**	**0.220**	**0.458**	**0.696**	**0.178**	**0.437**
Text-only	$Linear_1$	AudioCaps	0.609	0.423	0.286	0.181	0.204	0.429	0.602	0.174	0.388
Text-only	Gaussian	Clotho	0.524	0.339	0.222	0.136	**0.173**	0.371	0.379	0.132	0.256
Text-only	$Linear_1$	Clotho	**0.568**	**0.375**	**0.251**	**0.158**	0.172	**0.378**	**0.394**	0.127	**0.261**

Table 3: All models use AudioCaps and Clotho datasets in training. Symbol † indicates that LLM-generated text [18] is added in training.