Paper ID: 57

Automated Audio Captioning by Fine-tuning BART on AudioSet Tags

Félix Gontier, Romain Serizel and Christophe Cerisara
Université de Lorraine, CNRS, Inria, Loria

Context

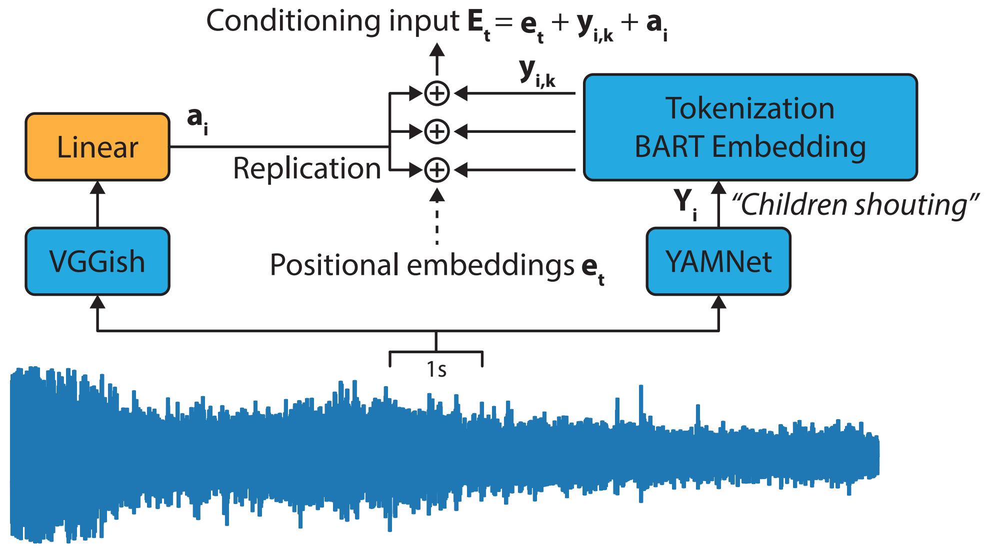
Task: Describe the contents of audio extracts with fluent English sentences
—> Contextualize sound events to eventually achieve a higher level of understanding of audio scenes

Current approaches:
- Encoder-decoder models with attention
  —> *Language modeling is learned from scratch*
- Textual conditioning provides vocabulary guidance in addition to audio embeddings
  —> *Audio-based keyword extraction or caption retrieval systems trained on the captioning dataset*

Proposition: Combine pre-trained audio tagging and large-scale language models

- https://github.com/felixgontier/dcase2021aac

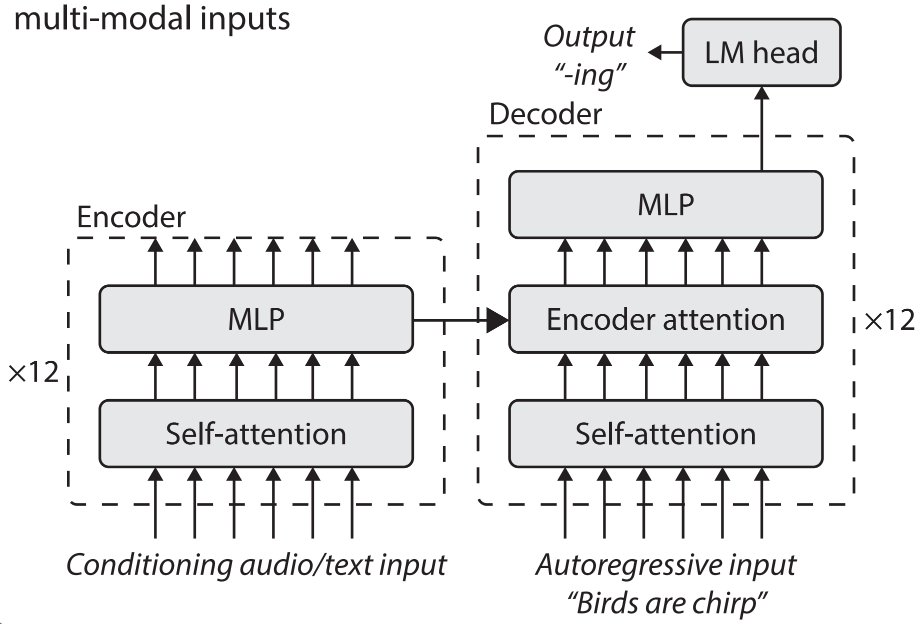
Multimodal Conditioning



- Closed vocabulary of 521 YAMNet (AudioSet) tags
- Single learnable linear layer
- Audio representation added to text and positional embeddings

Proposed Approach

BART conditional language model [1]
- Standard transformer encoder-decoder architecture
- Byte-pair encoding tokenization: 50265 tokens in vocabulary
- Pre-training scheme: denoising heavily corrupted text
- Transfer learning to AAC by simply fine-tuning with multi-modal inputs



Experimental Setup

AudioCaps dataset [2]
- Training on 49000 clips, 10s each, single caption
- Validation/Evaluation on 485/955 clips, 5 captions
- Subset of AudioSet: **in-domain audio** for pre-trained tagging models

Training:
- Cross-entropy loss on BART token vocabulary
- Stable training observed until convergence, even when fully fine-tuning
- Results reported over 3 runs with different random seeds
- Sampling on YAMNet logits at training: data augmentation
- Most likely tags taken at inference

Evaluation:
- Machine translation metrics (n-gram matching): BLEU-1/2/3/4, METEOR, ROUGE-L, **CIDEr**
- Graph parsing metric for semantics: **SPICE**
- **SPIDEr**: average of CIDEr and SPICE, main metric of DCASE task 6

Results

Conditioning setup
- Sequential embeddings are critical to captioning
- Text-only conditioning is better than audio-only
  —> Better use of the text-only BART pre-training setting?
- Both PANNs and VGGish complement YAMNet, PANNs are more informative

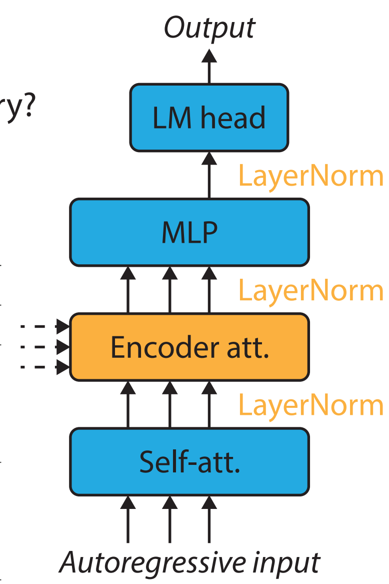
VGGish	PANNs	YAMNet	CIDEr	SPICE	SPIDEr
	×		6.5 (2.5)	6.1 (0.9)	6.3 (1.7)
×			37.6 (8.0)	11.9 (1.3)	24.7 (4.6)
		×	54.7 (0.6)	14.1 (0.2)	34.4 (0.4)
×		×	63.9 (1.0)	15.9 (0.3)	39.9 (0.7)
	×	×	**75.3 (0.9)**	**17.6 (0.3)**	**46.5 (0.6)**

Model performance
- On par or better than the state of the art on AudioCaps
- Higher BLEU-1/2/3 than reference captions cross-validation (human score)

Model	CIDEr	SPICE	SPIDEr
TopDown-AlignedAtt [2]	59.3	14.4	36.9
Koizumi et al. [3]	50.3	13.9	32.1
Eren et al. [4]	75.0	-	-
BART + YAMNet + PANNs	**75.3**	**17.6**	**46.5**
Human	91.3	21.6	56.5

Complementary experiments
- **Random initialization:** The performance improvement from BART pre-training is limited with sufficient amounts of training data
- **Freezing:** BART decoder is already effective to model caption structure
- **Capacity:** A marginal decrease in performance is observed with 3 times fewer parameters
  —> Low diversity in caption structure and vocabulary?
- **Task-specific fine-tuning:** The initial training loss is lower with summarization checkpoints

Variant	CIDEr	SPICE	SPIDEr
BART-large (400M)	75.3	17.6	46.5
No pre-training	71.0	16.7	43.8
Frozen decoder	68.5	16.6	42.5
BART-base (140M)	73.1	16.8	45.0
BART-XSum	71.8	17.3	44.5
BART-CNN	72.2	17.7	44.2
BART-CNN, frozen decoder	70.4	15.6	43.0



References:

[1] Lewis et al., BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. ACL 2020.
[2] Kim et al., AudioCaps: Generating captions for audio in the wild. NAACL 2019
[3] Koizumi et al., Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. 2020
[4] Eren et al., Audio captioning based on combined audio and semantic embeddings. ISM 2020

DCASE 2021 Workshop - 15-19 November 2021