Evaluating Off-the-Shelf Machine Listening and Natural Language Models for Automated Audio Captioning

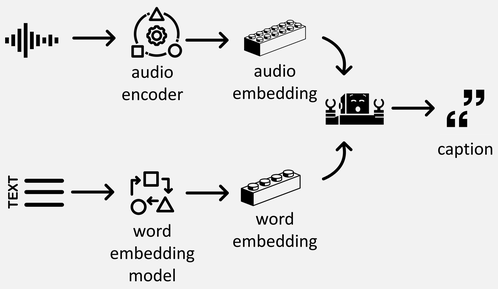
PRESENTER:
Benno Weck

BACKGROUND:

Find resources useful for Audio Captioning by comparing pre-trained models from Natural Language Processing and Machine Listening.

METHODS

1. Transformer-decoder as caption generator
2. Audio embeddings are fixed.
3. Word embeddings: fine-tuned & fixed.
4. Try different adapters on top of the audio embeddings and different hop sizes.
5. Test all combinations and compare by SPIDEr score.



RESULTS

- YAMNet is the best audio encoder in our comparison
- BERT is the best word embedding model in our comparison.
- Pre-trained word embeddings work better than randomly initialized.
- They work even better when fine-tuned!
- Extracting overlapping audio embeddings gives a small boost in performance.

Lego head icon: lego by Musket from the Noun Project
Icons in the flow-chart: process by Kamin Ginkaew from the Noun Project; process by Andrejs Kirma from the Noun Project; lego by jon trillana from the Noun Project; Robot by Drew Ellis from the Noun Project

YAMNet + BERT are the **best pre-trained** audio and language **models** for **audio captioning**.*

**conditions apply*

Pre-trained embeddings + Transformer =

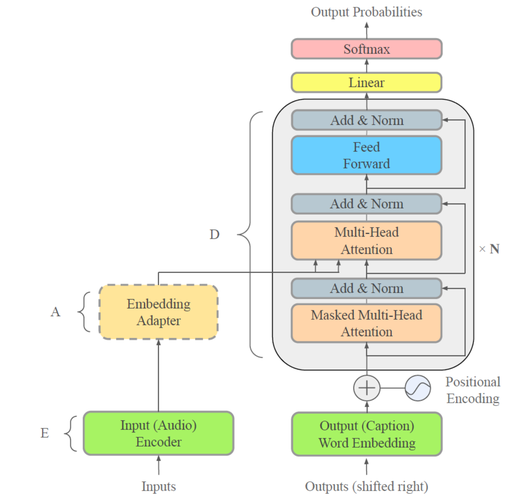
YAMNet + BERT =

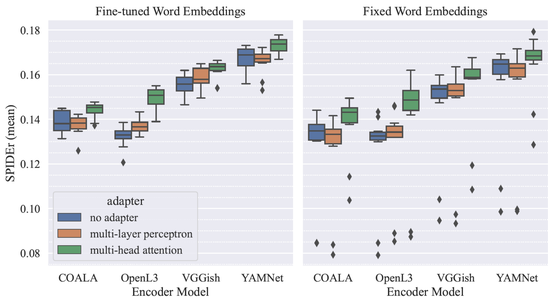
- YAMNet
- VGGish
- OpenL3
- COALA

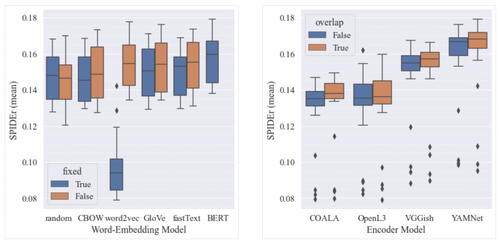
- BERT
- fastText
- GloVe
- word2vec
- randomly initialized

adapted
~~fine-tuned~~

fine-tune!







Encoder	Word embedding	Adapter	SPIDEr Mean	SPIDEr SD
COALA†	BERT	MHA-based	0.1495	0.0044
OpenL3⋆	BERT	MHA-based	0.1620	0.0051
VGGish⋆	BERT	MHA-based	0.1677	0.0052
YAMNet†	BERT	MHA-based	0.1793	0.0066

Benno Weck, Xavier Favory, Konstantinos Drossos, Xavier Serra

Universitat Pompeu Fabra Barcelona
HUAWEI
Tampere University

Lego Smile by TMD from the Noun Project