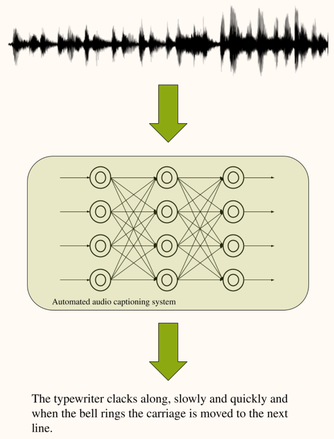
T6 Automated Audio Captioning

Coordinators

Samuel Lipping, Konstantinos Drossos, Tuomas Virtanen

Task description

- **Automatic generation of natural language descriptions from audio**
- Motivation: Automatic content description applications
- Example: Generation of audio descriptions for the auditorily impaired



Dataset

- Extended version of the AAC dataset from last year
- A total of 6974 15-30s audio files
- Five captions for each audio file, ranging from 8 to 20 words
- **Development split**
  - **Development-training split:** 3839 files, 24.0 hours
  - **Development-validation split:** 1045 files, 6.6 hours
  - **Development-testing split:** 1045 files, 6.5 hours
- **Evaluation split:** 1043 files, 6.6 hours

Submissions

- 37 Systems, 13 Teams, 52 Authors, 20 Affiliations
- Primarily **evaluated using SPIDEr**, a linear combination of the captioning metrics CIDEr and SPICE

Results, Top 10 teams

System	Audio encoder	Decoder	SPIDEr
Yuan	PANNs	Transformer	0.310
Xu	CNN	RNN	0.296
Xinhao	CNN	Transformer	0.294
Ye	ResNet38	RNN	0.280
Chen	ResNet38 + Memory Transformer	Meshed Transformer	0.262
Won	CNN	Transformer	0.249
Narisetty	Conformer + 1D/2D CNN	Transformer + RNN Language Model	0.236
Labbe	CNN	RNN	0.221
Liu	CNN	Transformer	0.184
Eren	1D/2D CNN + RNN	RNN	0.182
Baseline	RNN	RNN	0.012

Discussion

- Architectures
  - The most common encoder types were **CNNs** (33/37), followed by **transformers** (8/37), **RNNs** (3/37), and **MLP-mixers** (2/37)
    - Two types of transformer encoder: **Memory transformer** (4/37) and **Conformer** (4/37)
    - Transformer encoders were used together with CNNs and a CNN/RNN pair was also used
  - **Transformers** (23/37) and **RNNs** (14/37) were employed as decoders (23/37)
    - Two types of transformer decoder: **Regular** (22/37) and **meshed** (1/37)
- Learning setup
  - Most systems employed **transfer learning** for audio encoding, most commonly **AudioSet** (31/37) and **AudioCaps** (10/37), with the top 4 systems also using **crawled data**
    - Notably, all but one of the top 10 teams used **transfer learning with AudioSet**
  - All systems were trained with **supervised learning**, while a few also used **reinforcement learning** (4/37)
  - All systems used **cross-entropy loss**, one system also used a **sentence-level loss**
- Input data
  - Most systems used a **learned or pretrained word embedding** (32/37), while others used **one-hot word encoding** (5/37)
  - The top 28 systems relied on **data augmentation**, with some using **more than one** type of augmentation (8/37)