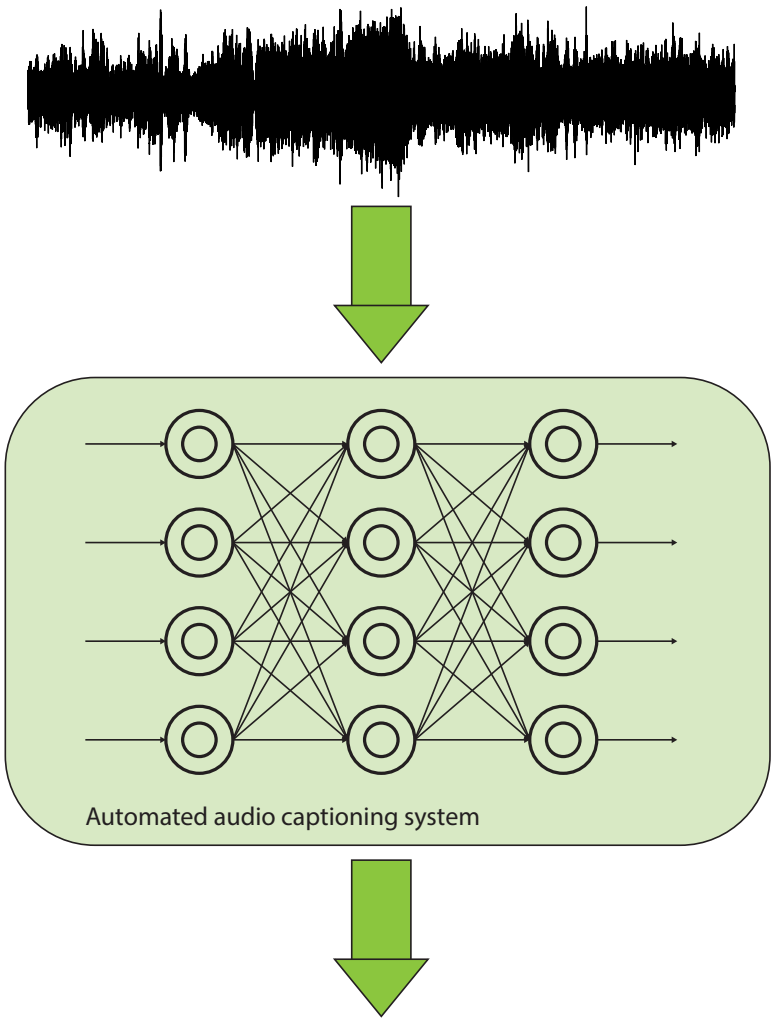
DCASE2022 Challenge task 6

Subtask A: Automated Audio Captioning

Félix Gontier, Huang Xie, Konstantinos Drossos
Samuel Lipping, Tuomas Virtanen, Romain Serizel

November 2, 2022

Task presentation



- **Describe acoustic scenes with natural language**
- Clotho dataset
  - Total of 6972 acoustic scenes from 15s to 30s
  - Five reference captions for each scene, 8-20 words

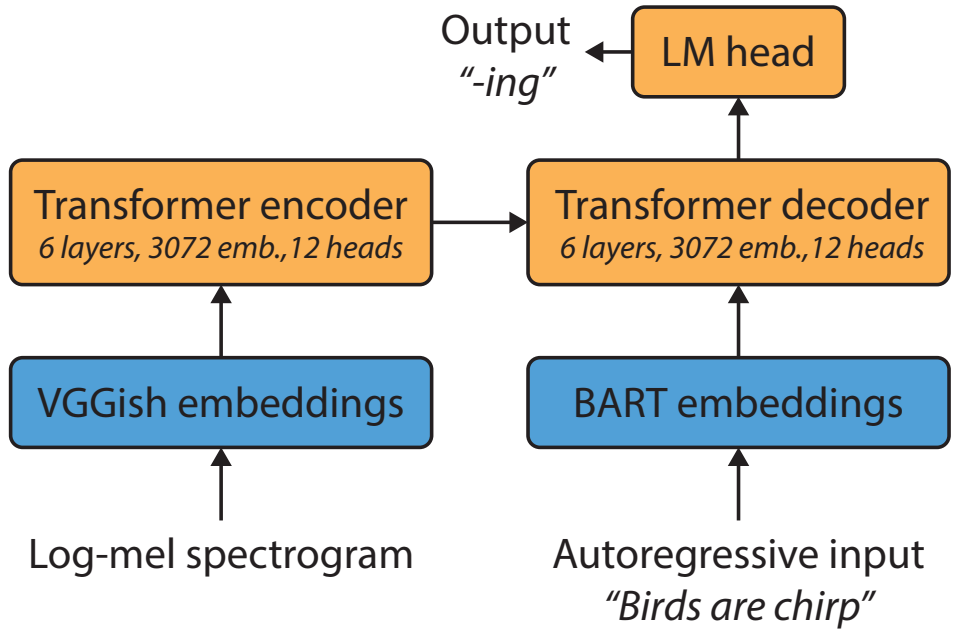
Subset	Split	# Files	Duration (h)
Development	Training	3839	24.0
	Validation	1045	6.6
	Evaluation	1045	6.5
Testing	-	1043	6.6

  - **New analysis subset** based on the Development-Evaluation split for robustness assessment (**microphone**, **scene mixing**, **environmental noise**)

Task presentation

Baseline system

- Transformer encoder-decoder
- Conditioned on frozen VGGish embeddings
- Pre-trained BART text embeddings
- Cross-entropy loss with teacher forcing



Evaluation

- Ranking metric: **SPIDEr** (combination of CIDEr and SPICE)
- Contrastive metrics: **Sentence-BERT** and **FENSE**

Task results

- Participation: 9 teams, 36 systems, 37 authors
- All teams managed to outperform the baseline
- Top 3 submissions used reinforcement learning to optimize CIDEr

Author	Audio encoder	Decoder	SPIDEr	Sentence-BERT	FENSE
Xu	PANNs (CNN14) + BiGRU	Transformer, GRU	**31.9**	50.8	22.7
Zou	PANNs (ResNet38)	LSTM	**31.8**	49.4	22.5
Mei	PANNs (CNN14)	Transformer	**30.9**	49.6	34.8
Primus	PANNs (CNN10) +Transformer	Transformer	**29.6**	47.6	44.0
Kouzelis	PaSST	Transformer	**29.3**	51.7	51.1
Guan	PANNs (CNN10)	Transformer (GraphAC, LocalAFT)	**29.1**	49.2	48.4
Kiciński	PANNs (CNN14)	Transformer	**27.0**	48.1	47.3
Pan	PANNs (CNN10)	Transformer	**25.5**	47.4	45.6
Labbé	PANNs (CNN10)	Transformer	**24.1**	47.5	45.2
Baseline	VGGish + Transformer	Transformer	**22.4**	45.4	44.6

System characteristics

General trends

- Log-mel spectrogram and pre-trained AudioSet model: **PANNs**
- Transformer architecture for language modeling
- Use of external data: **AudioCaps**, **MACS**
- Data augmentation: **SpecAugment**, **MixUp**
- Fitting evaluation metrics:
  - Validation monitoring
  - **Reinforcement learning on CIDEr**

Specific approaches

- Recent or novel architectures: GraphAC, LocalAFT
- Text conditioning with keyword retrieval or estimation
- Transfer learning from audio retrieval