AST: Audio Spectrogram Transformer

5 Apr 2021 · Yuan Gong, Yu-An Chung, James Glass · Edit social preview

In the past decade, convolutional neural networks (CNNs) have been widely adopted as the main building block for end-to-end audio classification models, which aim to learn a direct mapping from audio spectrograms to corresponding labels. To better capture long-range global context, a recent trend is to add a self-attention mechanism on top of the CNN, forming a CNN-attention hybrid model... read more

PDF Abstract

Code

Edit

YuanGongND/ast	official	★ 12	PyTorch

Tasks

Edit

Audio Classification · Audio Tagging · Classification · General Classification · Keyword Spotting

Datasets

Edit

AudioSet · Speech Commands · ESC-50

Results from the Paper

Edit

Ranked #1 on Keyword Spotting on Speech Commands (using extra training data)

Get a GitHub badge

Task	Dataset	Model	Metric Name	Metric Value	Global Rank	Uses Extra Training Data	Result	Benchmark
Audio Tagging	AudioSet	Audio Spectrogram Transformer	mean average precision	0.485	# 1	✓		Compare
Audio Classification	AudioSet	Audio Spectrogram Transformer	Test mAP	0.485	# 1	✓		Compare
Audio Classification	ESC-50	Audio Spectrogram Transformer	Top-1 Accuracy	95.7	# 1	✓		Compare
			PRE-TRAINING DATASET	AudioSet, ImageNet	# 1	✓		Compare
			Accuracy (5-fold)	95.7	# 1	✓		Compare
Keyword Spotting	Google Speech Commands	Audio Spectrogram Transformer	Google Speech Commands V2 35	98.11	# 1	✓		Compare

Keyword Spotting	Speech Commands	Audio Spectrogram Transformer	Google Speech Commands V2 35	98.11	# 1	✓		Compare

Methods

Edit

Adam • BPE • Dense Connections • Dropout • Label Smoothing • Layer Normalization • Multi-Head Attention • Residual Connection • Scaled Dot-Product Attention • Softmax • Transformer