Patchout Audio Transformers on Large-Scale Audio-Visual Data: Cross-Domain Transfer Performance

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Do Patchout Audio Transformers trained on large-scale audio-visual datasets outperform smaller-scale audio-only models in cross-domain tasks, as evaluated by transfer learning accuracy on AudioSet. The success of supervised deep learning methods is largely due to their ability to learn relevant features from raw data. Deep Neural Networks (DNNs) trained on large-scale datasets are capable of capturing a diverse set of features, and learning a representation that can. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Learning General Audio Representations with Large-Scale Training of Patchout Audio Transformers. Research question: Do Patchout Audio Transformers trained on large-scale audio-visual datasets outperform smaller-scale audio-only models in cross-domain tasks, as evaluated by transfer learning accuracy on AudioSet and UrbanSound8K benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

12 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The HEAR-eval tool was used to generate results for all PaSST models.	×	0.04
The official results of the HEAR 2021 challenge were produced using the HEAR-eval tool.	×	0.05
The HEAR 2021 tasks were divided into three categories: Speech, Music, and General.	×	0.06
The Speech category contains datasets consisting of verbal articulation of humans.	×	0.04
The Music category contains datasets consisting of instrumental sounds.	×	0.04
The General category contains mostly environmental sounds and acoustic scenes, but also others that cannot be clearly pu	×	0.03
Different metrics and ranges are used to evaluate the 19 tasks.	×	0.03
The maximum test score achieved by a model in the official HEAR 2021 challenge corresponds to 1 when tasks are grouped i	×	0.04
The baseline model of PaSST uses features M + H, a hop size of 10 ms, and a receptive field size of 160 ms.	×	0.07
DCASE 2016 Task 2 and MAESTRO rely on timestamp embeddings.	×	0.03
OpenL3 is trained in a self-supervised way using the audio and videos of Audioset.	×	0.05
Cat XWC is a concatenation of three diverse models: HuBERT, Wav2Vec2, and another model.	×	0.03
HuBERT is one of the models used in Cat XWC.	×	0.06

References

- http://arxiv.org/abs/2306.00830v1
- http://arxiv.org/abs/2505.07609v1
- http://arxiv.org/abs/2211.13956v2