Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford \* 1 Jong Wook Kim \* 1 Tao Xu 1 Greg Brockman 1 Christine McLeavey 1 Ilya Sutskever

**개요 (Abstract)**

We Study the capacity of speech processing systems trained simply to predict large amounts oftranscriptsof audio on the internet.

이 연구에서는 인터넷 상의 대량의 오디오 트랜스크립트를 예측하도록 훈련된 음성 처리 시스템의 기능을 연구합니다.

When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zeroshot transfer setting without the need for any finetuning.

680,000시간의 다국어 및 다중 작업 감독 하에서 훈련된 모델은 표준 벤치마크에서 잘 일반화되며, 추가 fine-tuning 없이도 이전의 완전 감독 결과와 경쟁할 수 있습니다.

When compared to humans, the models approach their accuracy and robustness.

인간과 비교했을 때, 이 모델은 정확성과 강건성 면에서 인간에 접근하고 있습니다.

We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

이 연구진은 강건한 음성 처리를 위한 기반으로 사용될 수 있는 모델과 추론 코드를 공개할 예정입니다

.

**1. 소개**

Progress in speech recognition has been energized by the development of unsupervised pre-training techniques exemplified by Wav2Vec 2.0 (Baevski et al., 2020).

음성 인식 분야의 발전은 Wav2Vec 2.0(Baevski et al., 2020)과 같은 비지도 사전 학습 기술의 개발로 인해 활기를 띠고 있습니다.

Since these methods learn directly from raw audio without the need for human labels, they can productively use large datasets of unlabeled speech and have been quickly scaled up to 1,000,000 hours of training data (Zhang et al., 2021), far more than the 1,000 or so hours typical of an academic supervised dataset.

이러한 방법은 사람의 레이블 없이 원시 오디오에서 직접 학습하므로, 레이블이 없는 대규모 음성 데이터를 효과적으로 활용할 수 있습니다. 이러한 접근법은 학술적 지도 학습 데이터셋의 일반적인 1,000시간보다 훨씬 많은 1,000,000시간의 데이터로 빠르게 확장되었습니다.

When fine-tuned on standard benchmarks, this approach has improved the state of the art, especially in a low-data setting.

표준 벤치마크에 파인튜닝하면 이 접근법은 특히 저 데이터 환경에서 최신 기술 수준을 향상시켰습니다,.