Tesorflow Speech Recognition Challenge

최 윤 철 (ycheol.choi@gmail.com)

Introduction

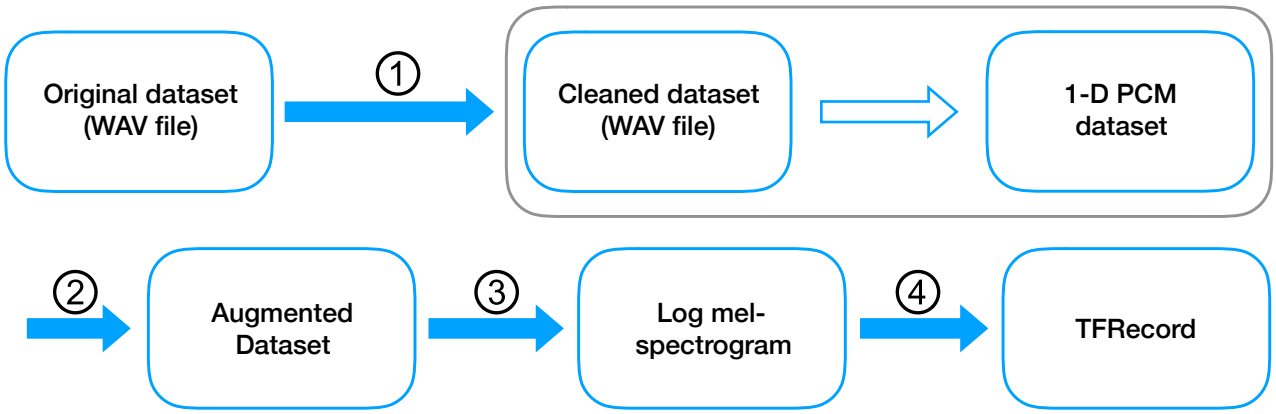
1초 길이의 오디오 데이터를 12가지 라벨 중 하나로 분류하는 문제 (Kaggle)

- **Training Data**
  - 라벨링 되어 있는 1초 길이의 오디오 파일 **64,728**개
  - 1분 길이의 background noise (6종류)
- **Test Data**
  - 라벨링 되어 있지 않은 1초 길이의 오디오 파일 **158,538**개

개발환경

- 서버 : AWS EC2 p3.2xlarge
  - OS: Ubuntu 16.04.3
  - GPU: 1 (16G)
  - CPU: 8
  - Memory: 61G,
- 언어 : Python 3.6.3
- 주요 라이브러리 : Tensorflow 1.4, librosa 0.5.1, Numpy 1.13.3, Pandas 0.21.0

Data Preprocessing



① Removing Abnormal Data

- Training Data에서 라벨과 일치하지 않거나 노이즈인 데이터를 제거
- 데이터를 spectrogram으로 변환한 뒤, 육안으로 데이터 이상 여부를 확인

② Data Augmentation

- 데이터 용량 및 학습시간을 고려해 원본데이터의 8배 크기로 augmentation
- **Speed tuning** : 원본 속도의 0.9 ~ 1.1배 범위에서 랜덤하게 속도 변경
- **Pitch tuning** : 한 옥타브를 24단계로 나누고 원본의 ±4단계 범위에서 랜덤하게 피치를 변경
- **Background noise mixing** : 6종류의 background noise 중에서 하나를 임의로 선택하여 원본 데이터와 합성

③ Making log mel-spectrograms

- Augmentation된 데이터셋을 log mel-spectrogram으로 변환
- librosa.feature.melspectrogram 메소드 사용

④ Converting dataset into TFRecord

- 학습시간 단축과 편의를 위해 ndarray 데이터를 TFRecord로 저장