Audio Clip Tagging

Everyday Sound Classification with CNN and Transfer Learning

Boyang Zhang
June 3 2019

Background

- Extensive research has been done before deep learning
  - Preprocessing methods
- Similarity between audio and image classification
  - Robust models for transfer learning
- Multiple Applications:
  - Voice/Sound recognition
  - Noisy environment information extraction

Dataset

Based on Kaggle Competition *Freesound Audio Tagging 2019*

Two sources for data:

- Noisy web audio data from Flickr videos taken from the YFCC dataset (Yahoo Flickr Creative Common)
- Curated (manually tagged) audio clips from FSD (Freesound Dataset)

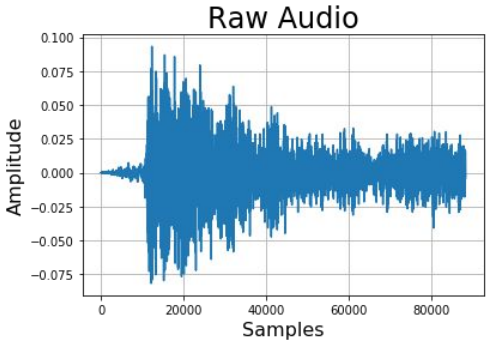
Properties:

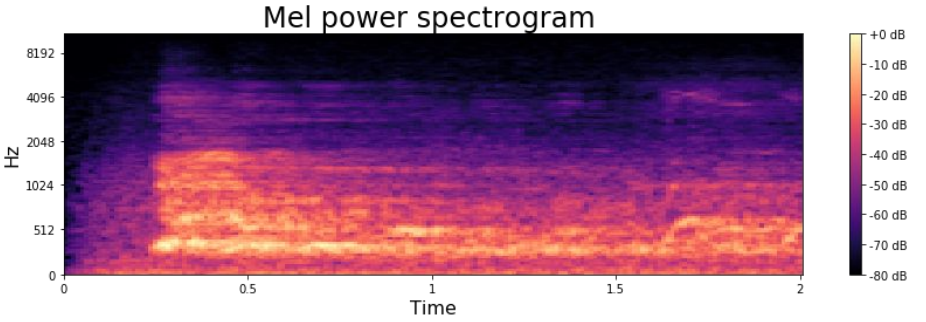
- 19815 noisy clips, 4970 curated clips
- Different length (0.2 to 30s)
- High sampling rates (44.1kHz)
- Large frequency range (10-22kHz)

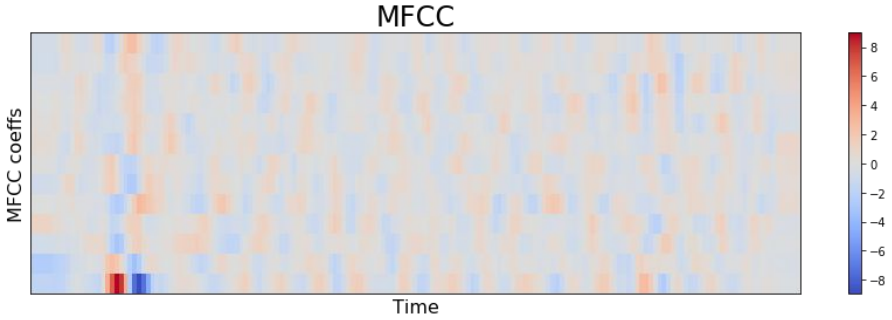
Data Preprocessing

- Trimming
- Padding
- Mel-spectrogram (filter banks)
- MFCC

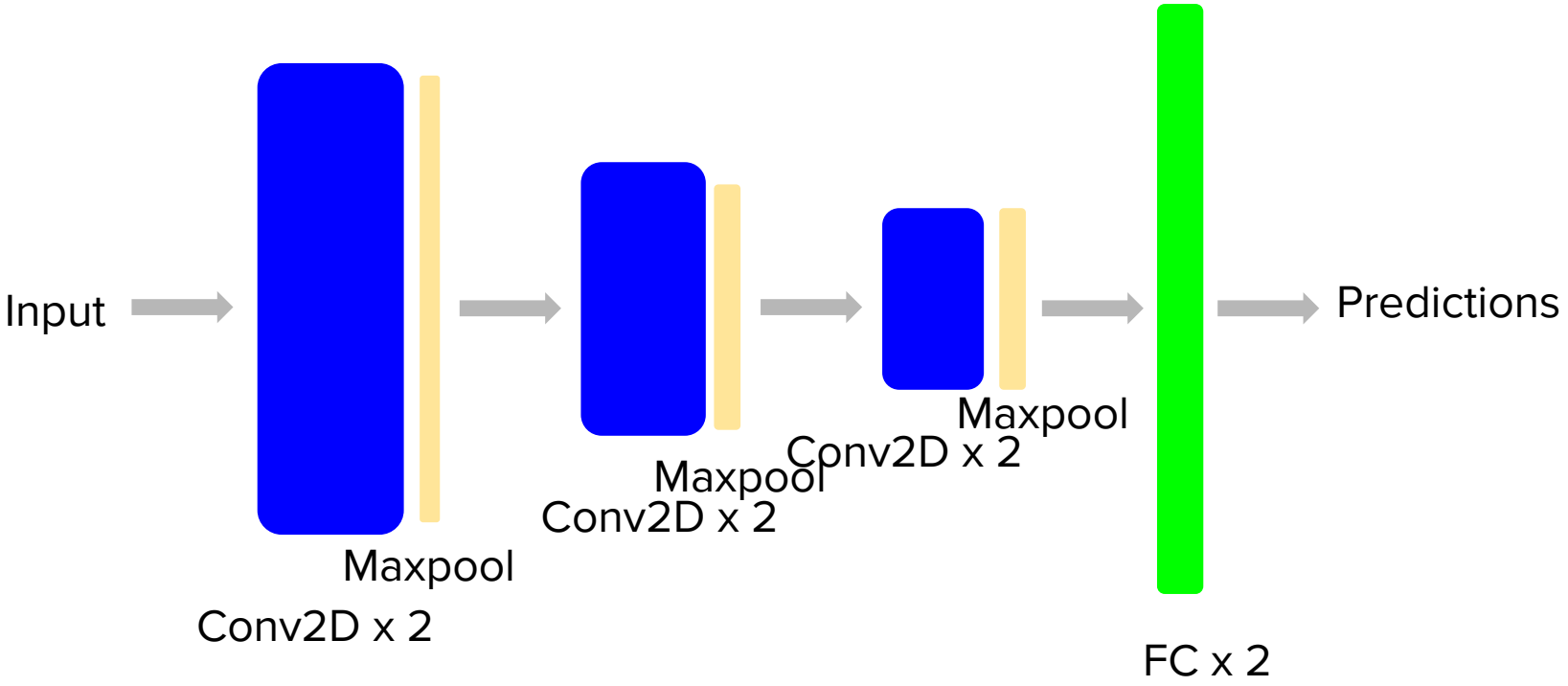
Example: Barking







CNN - architecture



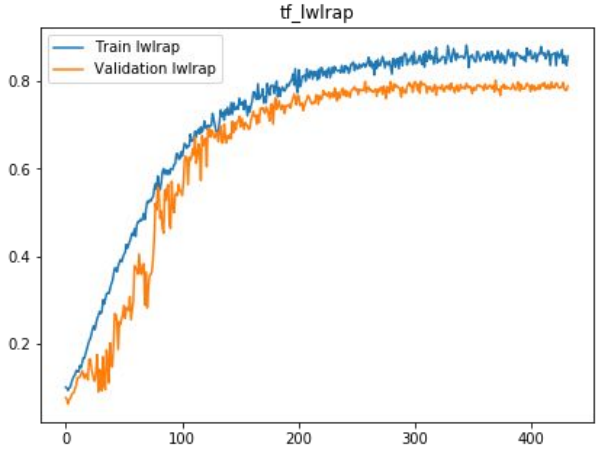
CNN - results

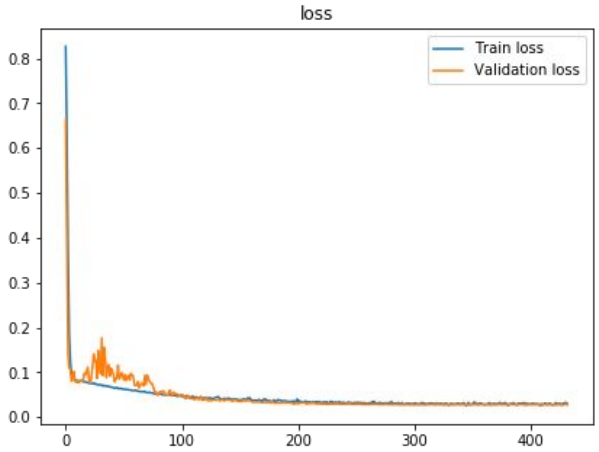
400+ Epochs

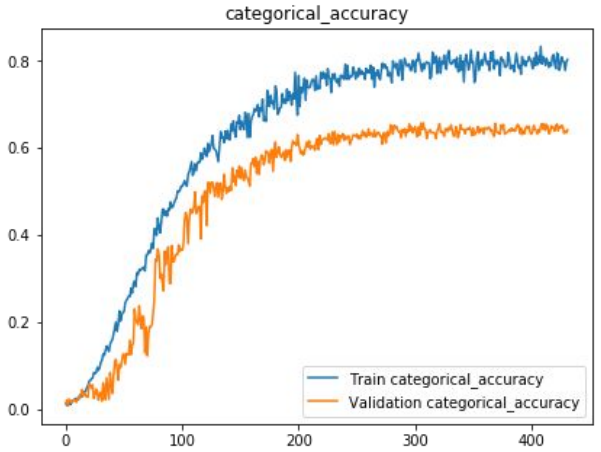
Categorical accuracy: 0.62; Score: 0.731

Competition Leaderboard:

Max:0.762 mean: 0.548 median: 0.61

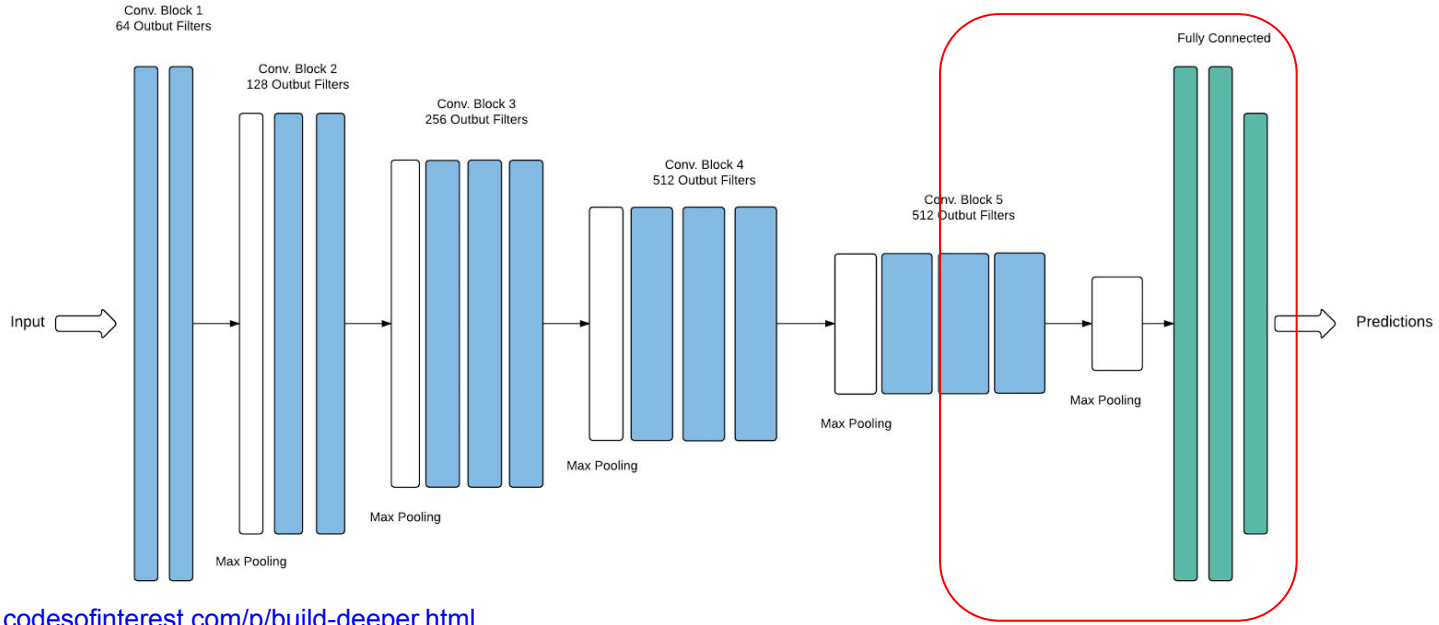






Transfer Learning

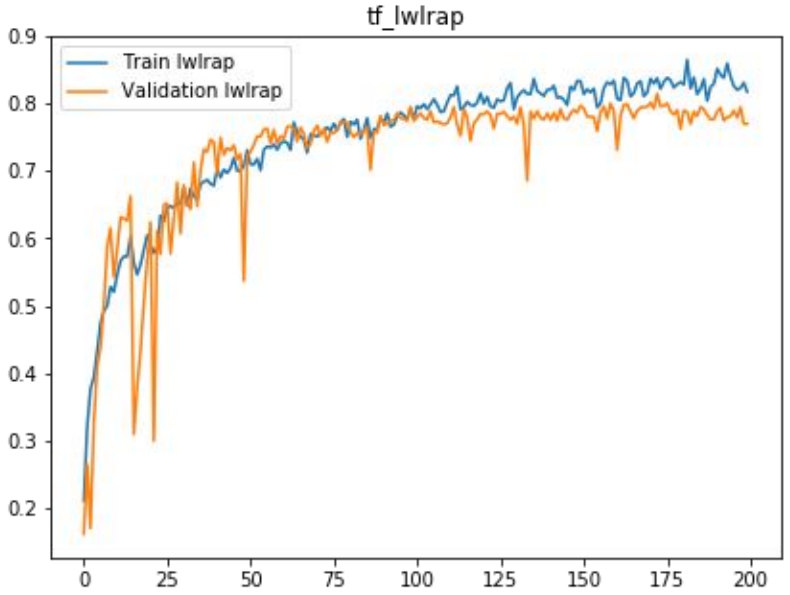
- VGG19 trained on ImageNet (Image classification model)
- Train customed fully connected layers only

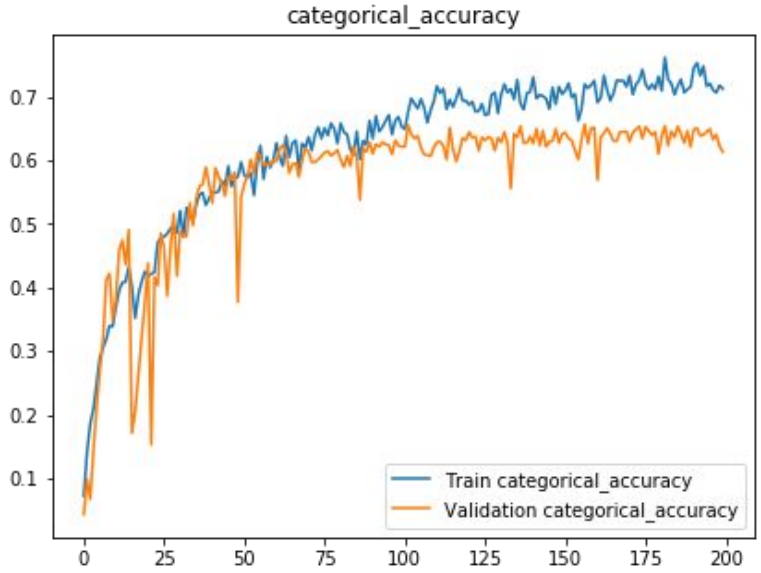


https://www.codesofinterest.com/p/build-deeper.html

Transfer Learning - results

200 Epochs; Categorical accuracy: 0.65; Score: 0.79





Doodle Recognition

Image Classification with Sequential Models

Background

- Abundant deep learning CNN based models for image recognition
- Sequential properties of doodle image data
- Handwriting analysis applications

Dataset and Preprocessing

Dataset:
Google AI “Quick! Draw” Dataset

340 categories 50M drawings

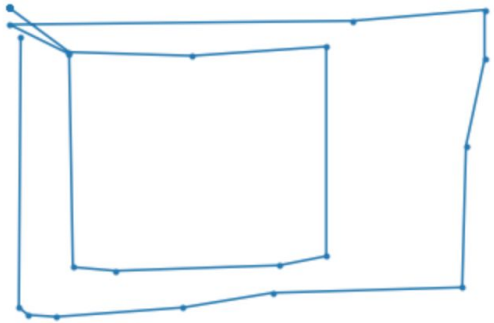
(For this project, 32 categories, 3M drawings)

Preprocessing:

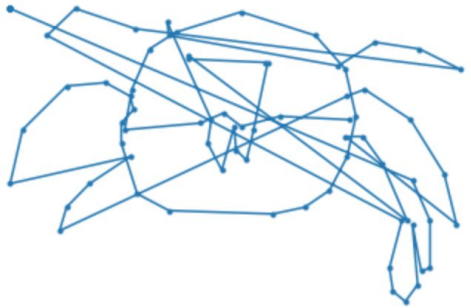
Strokes -> Sequences

Connect without losing information

Example: Microwave

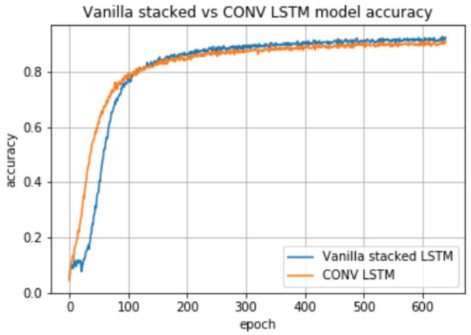


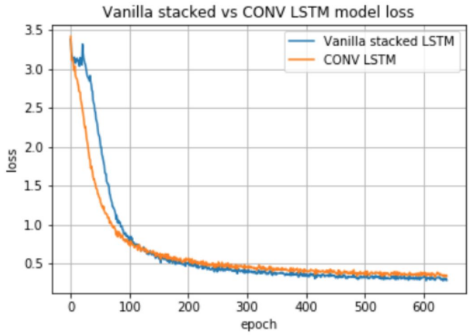
Example: Crab



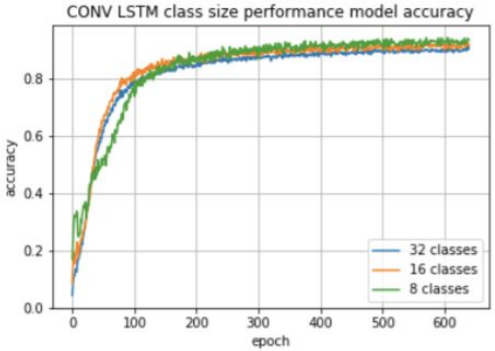
LSTM Structures

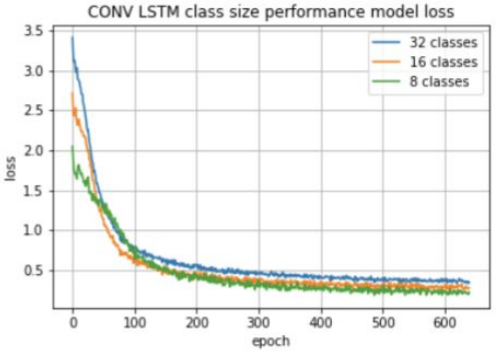
- Vanilla stacked LSTM
- Conv1D LSTM
- Performance under different class sizes (scalability)





32 Class Vanilla LSTM vs. Conv LSTM

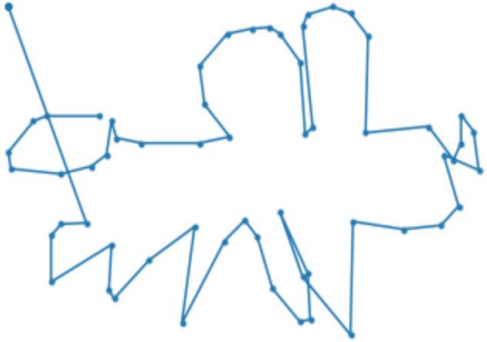




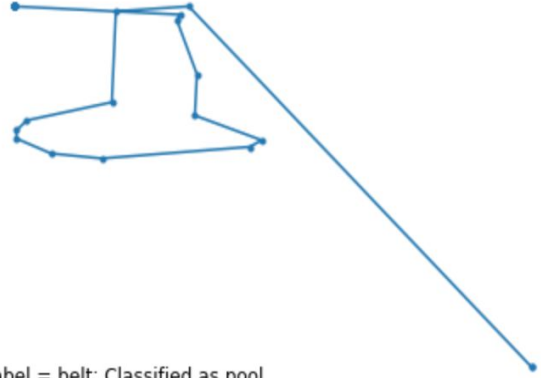
Conv LSTM Performance on Variable Amounts of Classes.

LSTM Structures - Examples

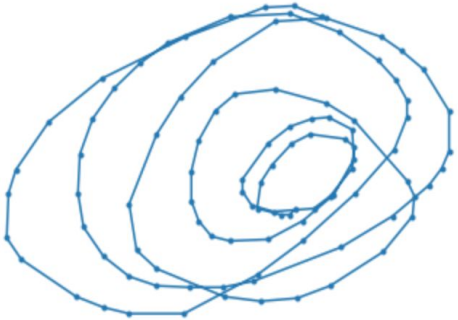
True/Classified label = camel



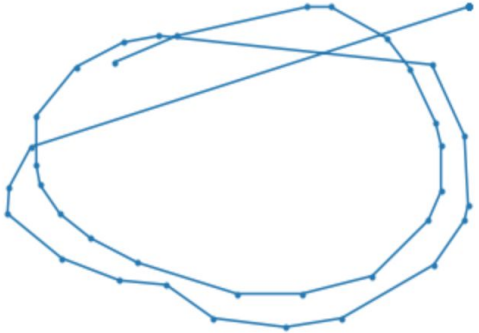
True label = crown; Classified as drill



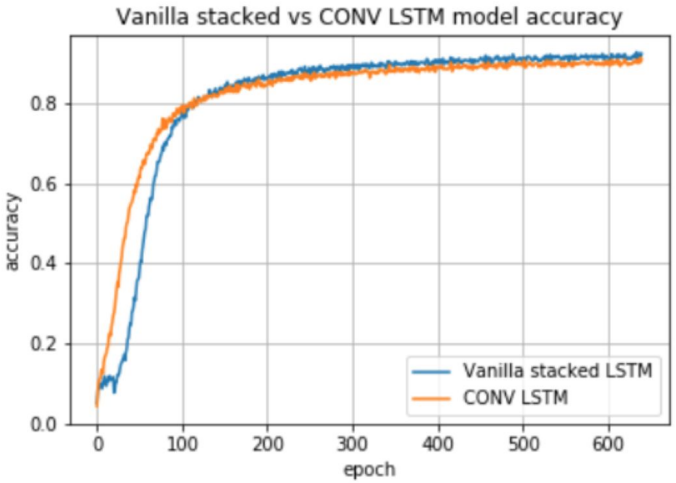
True/Classified label = onion

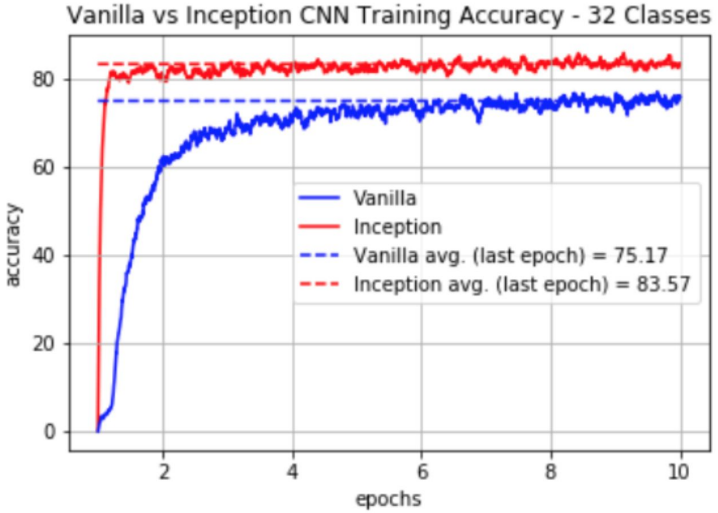


True label = belt; Classified as pool



LSTM vs CNN





	Vanilla CNN	Inception CNN	Vanilla LSTM	Conv LSTM
8 Classes	86.7%	92.5%	93.9%	94.2%
16 Classes	78.4%	86.8%	91.9%	91.4%
32 Classes	75.2%	83.6%	91.8%	90.9%