# Wave2Spec2Text

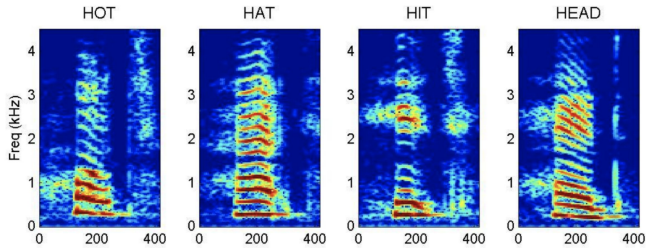
KAGGLE CHALLENGE

# Content

- **Challenge Introduction**
  - Dataset Description
  - Challenge with Speech Recognition
- **Implementation**
  - From Wave to Images
- **Experiments**
- **Conclusions**

# Challenge Description

- **Dataset:** TensorFlow Speech Commands Datasets of 65,000 one-second long utterances of 30 short words by thousands of different people
- **Dataset Labels:** “*yes*, *no*, *up*, *down*, *left*, *right*, *on*, *off*, *stop*, *go*”
- **Objective:** Build best classification model



# Challenge with Speech Recognition

United States

Water

(wootaa)

www.youtube.com/JapaneseEng101

England

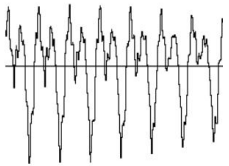
Australia

# Challenge with Speech Recognition

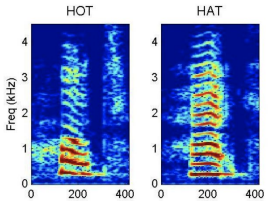
# Implementation

PYTORCH

- **From Wave to Images:**
  - Converting wave-format into spectrograms
  - Storing frequency (40 log-mel) and time in an image
  - Use padding to ensure the same dimensions



Speech

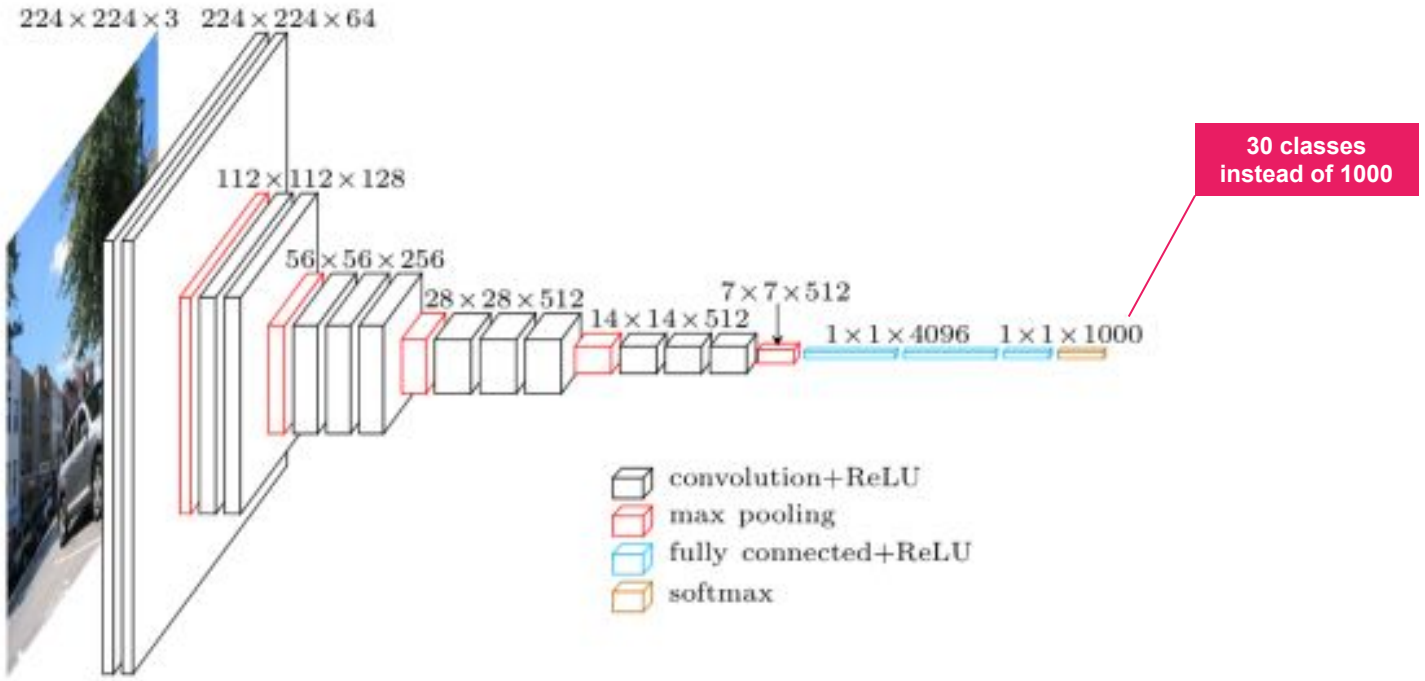



Spectrogram

“House”

Word

# VGG11 Architecture

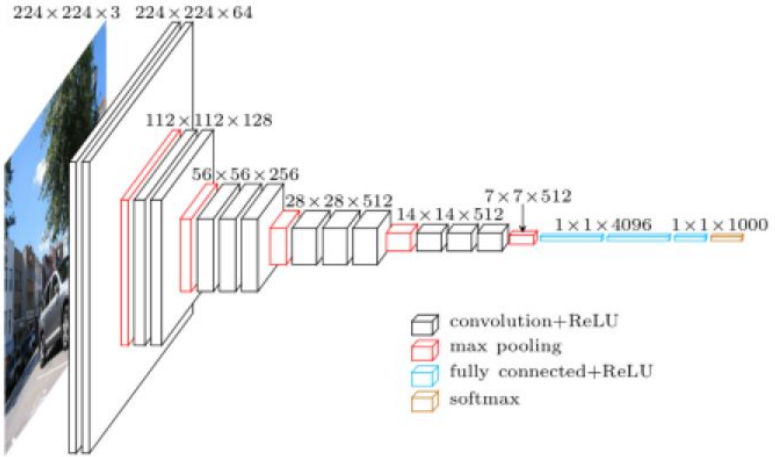


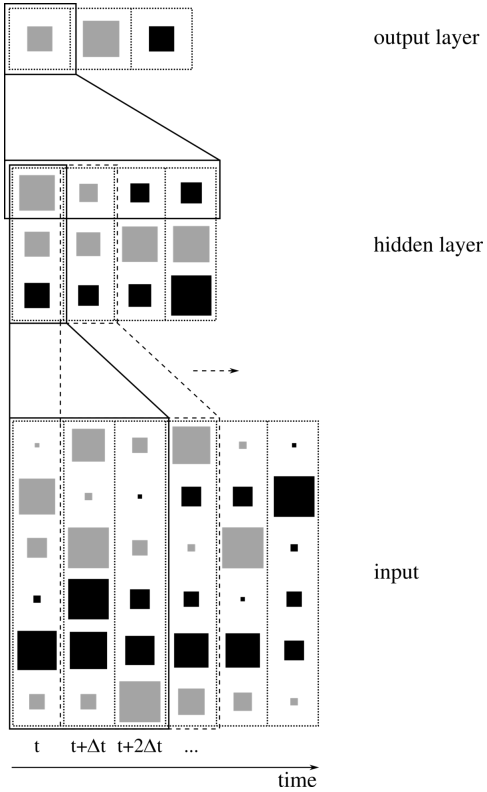
# List of Experiments

1. **Testing different Network Depths and Architectures (VGG)**
    - Will Deeper Networks Retain More Important Information?
        - Risk over-fitting (a common problem)
    - Will different architectures perform better?
2. **Testing Different Optimization Functions**
    - SGD vs. Adam
3. **Training with Augmented data**
4. **Using an Ensemble Model**
5. **Visualizing Confusion Matrix**

# Experiments 1: Testing Network Depths and Architectures

- TDNN, VGG11, VGG16
- 5 Epochs
- Checking for overfitting



# Experiment 1: Testing Network Depths and architectures

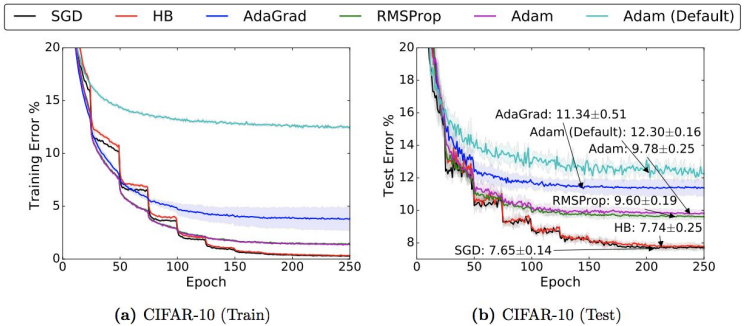
| Accuracy | TDD | VGG11 SGD | VGG11 | VGG16 | VGG16 * | Ensemble | Ensemble ** |
|---|---|---|---|---|---|---|---|
| **Validation** | 89.95% | 90.8% | 95.5% | 91.13% | 94.76% | N/A | N/A |
| **Test** | 89.53% | 90.4% | 93.47% | 90.84% | 94.7% | 95.19% | 95.54% |

VGG11 > VGG16!

* Trained on augmented data ** Weighted Ensemble

# Experiment 2: Optimization Functions

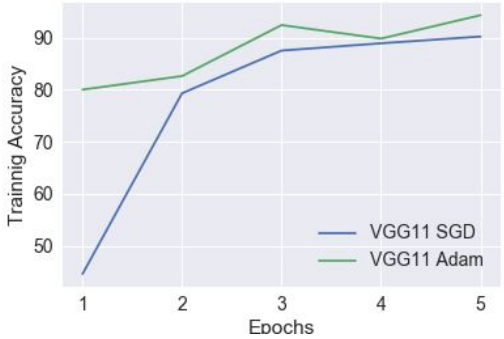
- From UC Berkeley paper *The Marginal Value of Adaptive Gradient Methods in Machine Learning*:
  - Adaptive models usually perform better during training (converge faster) than non-adaptive optimizers, but can generalize worse (CIFAR-10)




(a) CIFAR-10 (Train) (b) CIFAR-10 (Test)

- Optimizers for Speech2Spec2Text

| | A | B | C |
|---|---|---|---|
| 1 | | VGG11 Adam | VGG11 SGD |
| 2 | Validation Accuracy | 94.3% | 90.80% |
| 3 | Test Accuracy | 94.5% | 90.40% |




Source: https://arxiv.org/pdf/1705.08292.pdf, UC Berkeley & Toyota Technological Institute at Chicago, May 24, 2017

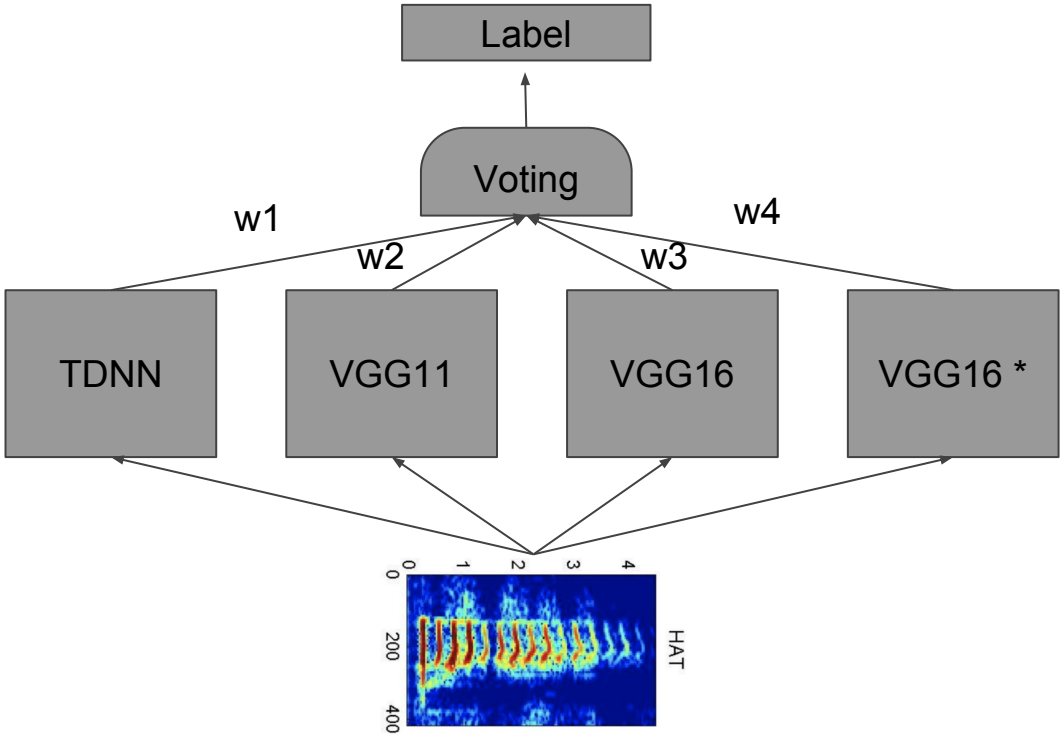
# Experiment 3: Data Augmentation

- Augmented each sample to 0.9 and 1.1x speed. 150,000 sample recordings.

| Accuracy | TDD | VGG11 SGD | VGG11 | VGG16 | VGG16 * | Ensemble | Ensemble ** |
|---|---|---|---|---|---|---|---|
| **Validation** | 89.95% | 90.8% | 95.5% | 91.13% | 94.76% | N/A | N/A |
| **Test** | 89.53% | 90.4% | 93.47% | 90.84% | 94.7% | 95.19% | 95.54% |

3.8% increase in Accuracy

# Experiment 4: Ensemble model

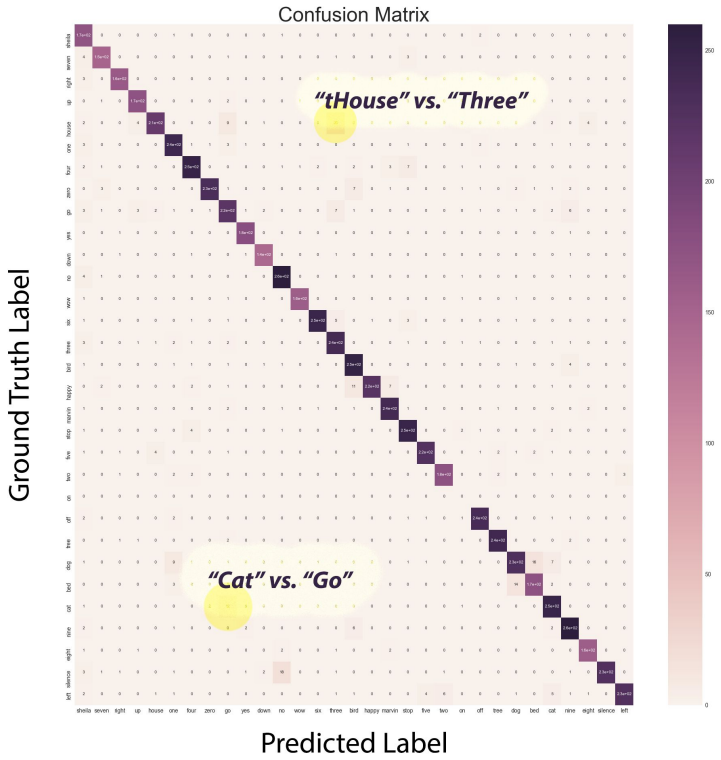



| Accuracy | Ensemble | Ensemble ** |
|---|---|---|
| **Validation** | N/A | N/A |
| **Test** | 95.19% | 95.54% |

** Weighted Ensemble

# Experiments 4: Confusion Matrix

Analysis of VGG11 SGD Model



# Conclusion

- Good but unrealistic results: 95.54% is state of the art - top 1 in Kaggle private dataset is 91.06% accurate.
- Several ways to go forward:
  - Further data augmentation
  - More models in ensemble
  - Explore other architectures
  - Exploiting domain specific knowledge
  - Analyze image embeddings together with confusions matrix