# Deep Neural Networks in Speech and Vision Systems

Published on May, 2020

Shailesha Prasad Maganahalli

# Introduction

- Neural Networks has mostly been limited to the application of industrial control and robotics.
- Recent advancement in Neural Networks have led to successful applications in Intelligent System.
- This presentation is intended to understand comprehensive surveys on the latest developments in intelligent vision and speech applications from the perspectives of both software and hardware systems.
- Many of these emerging technologies using deep neural networks show tremendous promise to revolutionize research and development for future vision and speech systems.
- The advancement in speech and vision processing systems has enabled tremendous research and development in the areas of human-computer interactions , biometric applications , security and surveillance , and most recently in computational behavioral analysis .
- Google, Facebook, Apple, Microsoft, IBM, and others have adopted deep learning as one of their core areas of research in artificial intelligence (AI).

# DESIGN AND ARCHITECTURE OF NEURAL NETWORKS FOR DEEP LEARNING

Neural network is designed similar way of human eye process visual information in hierarchical fashion. This hierarchical arrangement has more abstract to learn meaningful information at different levels. Such neural networks are termed as Artificial Neural Networks (ANN), also known as Deep Neural Networks (DNN) in literature.

Below list of notable NN models extract both simple and complex features similar hierarchical regions of the primate vision system.

1. Convolution Neural Networks (CNN)
2. Deep Belief Networks (DNN)
3. Stacked Auto-Encoders (SAE)
4. Generative Adversarial Networks (GAN)
5. Variational Auto Encoder (VAE)
6. Recurrent Neural Networks (RNN

# Deep learning in computer vision

Data Set : MIST database.

CNN has showed significant performance improvement in hand-written digit recognition task compared to earlier state-of-the-art machine learning techniques.

SUMMARY OF THE SIGNIFICANT STATE-OF-THE-ART CNN IMAGE CLASSIFICATION RESULTS
(*ACTUAL CLASS ERROR WITHIN TOP 5 PREDICTIONS, **PIXEL CLASS ERROR)

| Architecture | Dataset | Error rate |
|---|---|---|
| AlexNet [33] - University of Toronto 2012 | Imagenet (natural images) | 17.0%* |
| GoogLeNet [93] - Google 2014 | Imagenet (natural images) | 6.67%* |
| ResNet [71] - Microsoft 2015 | Imagenet (natural images) | 4.70%* |
| Squeeze & Excitation [100]– Oxford 2018 | Imagenet (natural images) | 2.25%* |
| Multiscale CNN [92] - Farabet et al. 2013 | SIFT/Barcelona (scene labeling) | 32.20%** |

# Deep learning in speech recognition

Data Set : TIMIT, Bing-Voice-Search speech, Switchboard speech, Google Voice Input speech, YouTube speech, and the English-Broadcast-News speech dataset.

DBN has shown significant improvement in performance. DBN has outperformed other speech recognition such as Gaussian mixture model (GMM)-HMM. SAEs likewise are shown to outperform (GMM)-HMM on Cantonese and other speech recognition tasks

SUMMARY OF THE SIGNIFICANT STATE-OF-THE-ART DNN SPEECH RECOGNITION MODELS (*PERPEPLEXITY-SIZE OF MODEL NEEDED FOR OPTIMAL NEXT WORD PREDICTION WITH 10K CLASSES, **WORD ERROR RATE)

| Architecture | Dataset | Error rate |
|---|---|---|
| RNN [126] - FIT, Czech Republic, Johns Hopkins University, 2011 | Penn Corpus (natural language modeling) | 123* |
| Autoencoder/DBN [128] - Collaboration, 2012 | English Broadcast News Speech Corpora (spoken word recognition) | 15.5%** |
| LSTM [129] - Google, 2014 | Google Voice Search Task (spoken word recognition) | 10.7%** |
| Deep LSTM [130] - National Chiao Tung University, 2016 | CHiME 3 Challenge (spoken word recognition) | 8.1%** |
| CNN-BLSTM [131] - Microsoft, 2017 | Switchboard (spoken word recognition) | 5.1% |
| Attention (LAS) & LSTM [132] - Google, 2018 | In-house google dictation (spoken word recognition) | 4.1% |
| Attention & LSTM with pretraining [133] - Collaboration, 2018 | LibriSpeech (spoken word recognition) | 3.54% |

# LIMITATIONS OF DEEP COMPUTATIONAL MODELS

Deep Learning progress in speech and visual applications might impact future progress due to below limitations.

1. The first area is to develop a robust learning algorithm for deep models that requires a minimal amount of training samples.
2. Effect of sample size
3. Computational burden on mobile platforms
4. Interpretability of models
5. Pitfalls of over-optimism

# Future and Upcoming Improvements

1. Implementing standalone vision and speech applications on mobile and resource constrained devices.

2. sophisticated deep learning-based intelligent systems. From sentiment and emotion recognition to developing self-driving intelligent transportation systems.

3. There are enormous opportunity in speech and visual applications which assist humans visual and speech perception to larger scale with precision.

# Thank You