Google’s DeepMind AI can lip-read TV shows better than a pro

Michael Wu

What is DeepMind?

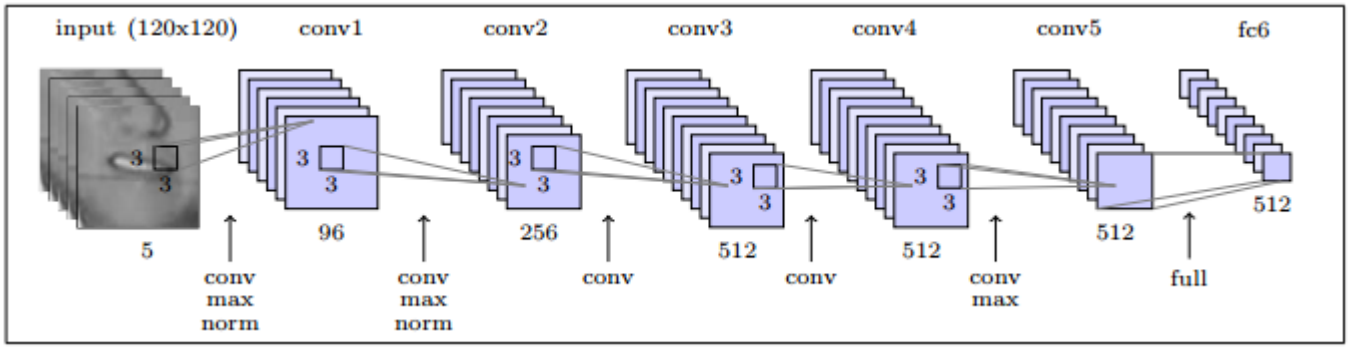
- A company owned by Alphabet, Google's parent company, that does AI research
- Uses neural networks, which are based on the human brain
- Many applications, will be focusing on speech recognition

The Goal

- Be able to recognize speech from video
- Very challenging for even humans to do
- Accuracy and versatility
- Can be used to dictate input to devices
- Used in the open world
- How to do this?

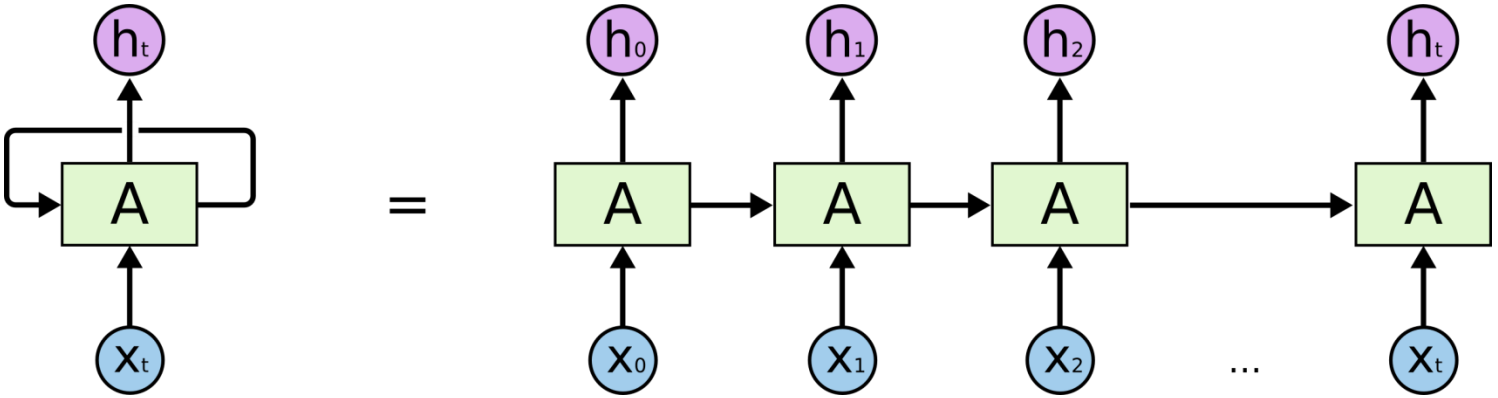
Neural Networks

- Convolutional Neural Network (CNN)
- Used for image processing
- Each neuron has a receptive field
- Deep learning with multiple layers
- Assigns weight to different filters



Neural Networks

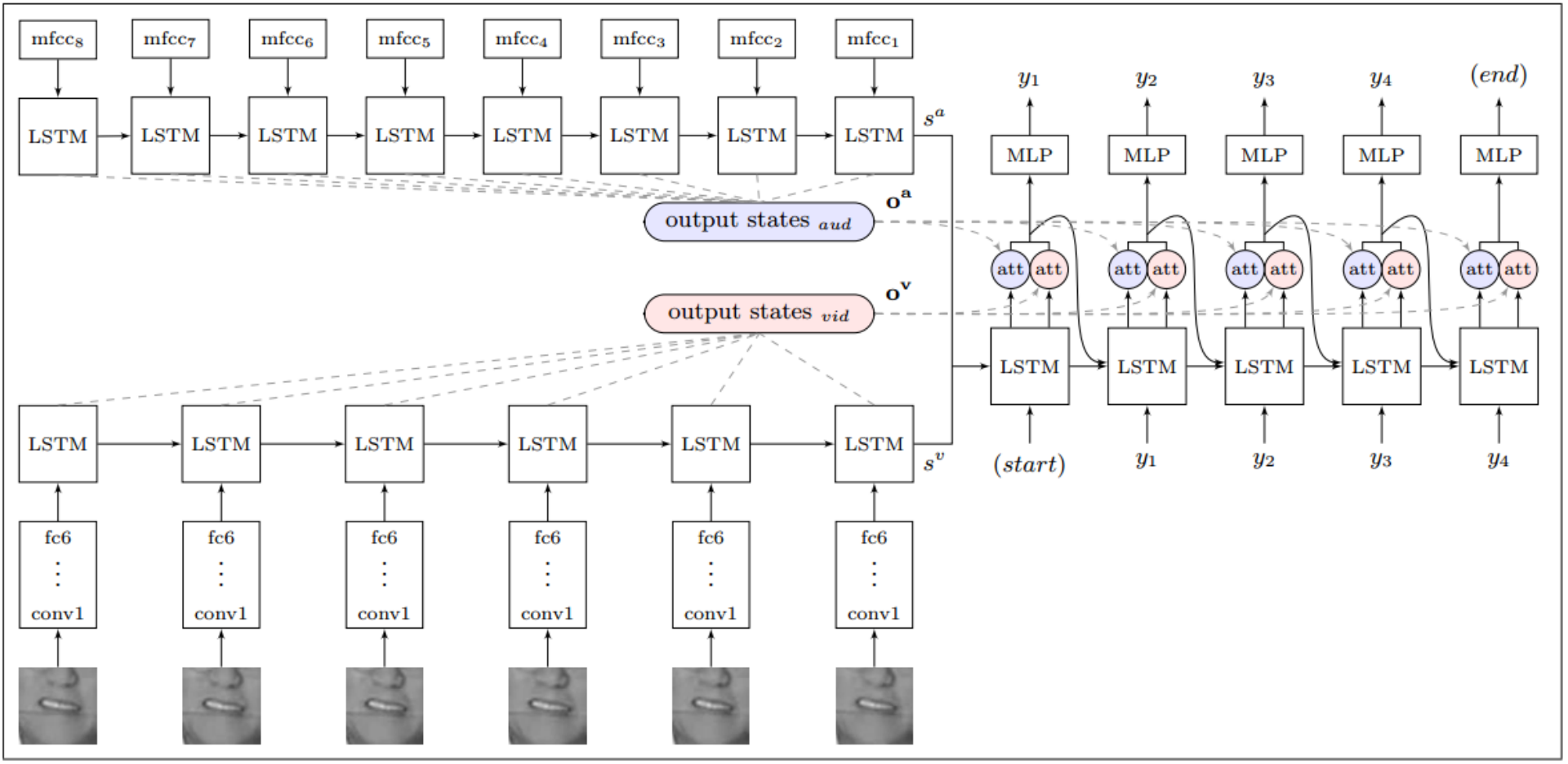
- Long Short-Term Memory (LSTM) network
- Type of Recurrent Neural Network (RNN)
- Normally networks operate independently of previous outputs
- LSTM networks allows previous data to be used



Watch, Listen, Attend, and Spell

- Two modules *Watch* and *Listen* to process inputs
- Both generate attention vectors used in the *Attend* module
- *Spell* takes the outputs of *Watch*, *Listen,* and *Attend,* to generate a probability distribution for characters based on previous data
- Multilayer perceptron (MLP)

Watch, Listen, Attend, and Spell



Training

- Generates the filters for the neural network
- Supervised training
- Runs data through the AI system and adjusts filters based on the correct output
- Needs large amount of data
- Audio, Visual, and Audio-Visual data prevents one channel dominating

Data

- Generated with facial recognition and audio subtitle alignment
- 4960 hours of video data from BBC
- Audio-only data set
- Significant improvement from other public data sets (GRID, LRW)

Data

- GRID consists of limited vocabulary (51 words)
- Recorded in controlled lab environment
- Lip Reading in the Wild (LRW) consists of 500 individual words from BBC broadcasts

Results

- On the BBC data 53.2% Word Error Rate using *Watch* only, compared to 73.8% Word Error Rate from a professional lip reader.

Methods	LRW [9]	GRID [11]
Lan *et al.* [23]	-	35.0%
Wand *et al.* [39]	-	20.4%
Chung and Zisserman [9]	38.9%	-
WAS (ours)	**15.5%**	**3.3%**

Takeaways

- DeepMind has created a model for lip reading that performs far better than previous efforts
- Eventually will be able to use this technology in consumer products
- DeepMind generated a data set for learning for future lip reading efforts

Citations

1. Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2016) Lip Reading Sentences in the Wild. *Computing Research Repository.* Retrieved from https://arxiv.org/list/cs.CV/1611.
2. Sak, H., Senior, A., Beaufays, F. (2014) Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *Computing Research Repository.* Retrieved from https://arxiv.org/list/cs.NE/1402.
3. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems, 25,* 1097-1105. Retrieved from https://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012.