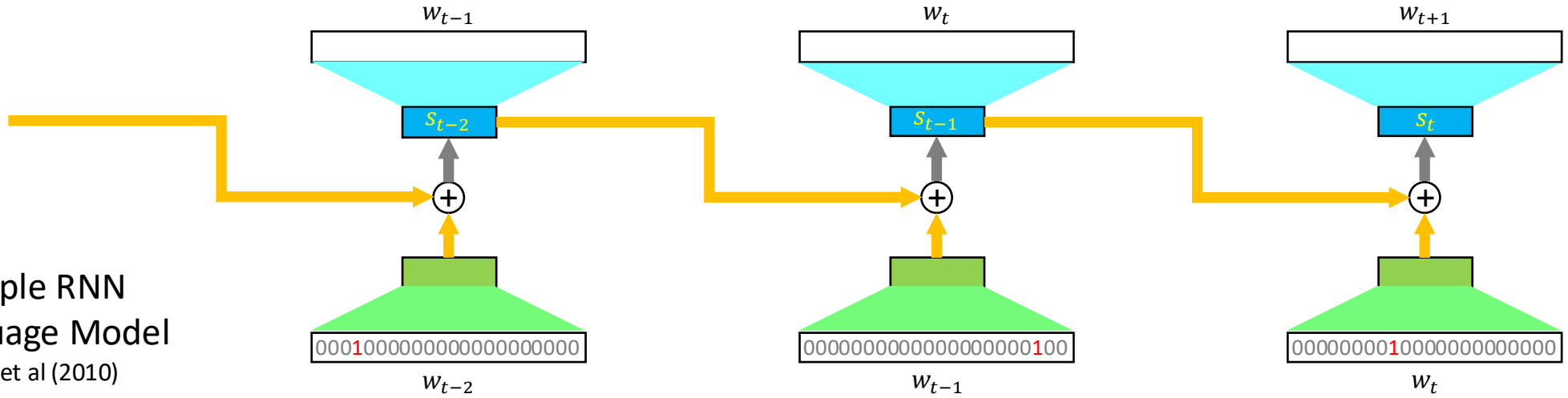
Self-Supervised Representation Learning for Automatic Speech Recognition

SSL for Speech using Masked Language Model Objective: Hsu et al (2021)
SSL for Speech using Noise Contrastive Objective: Schneider et al (2019), Baevski et al (2020)
Interpreting SSL as Maximum Mutual Information Estimation

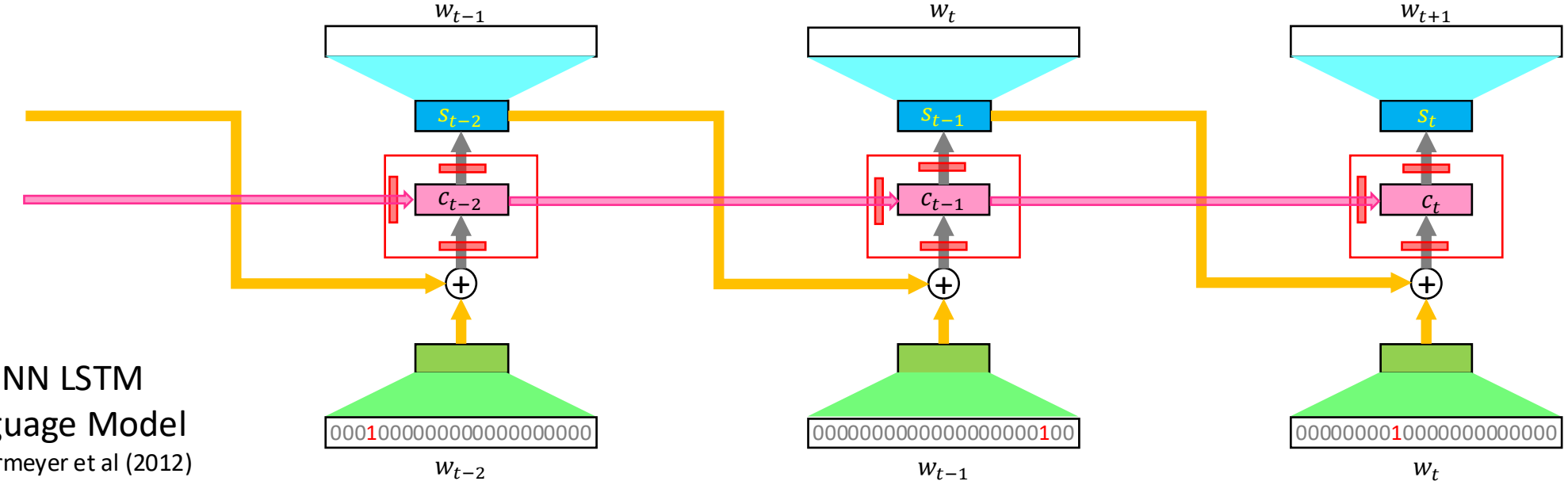
A Simple RNN Language Model

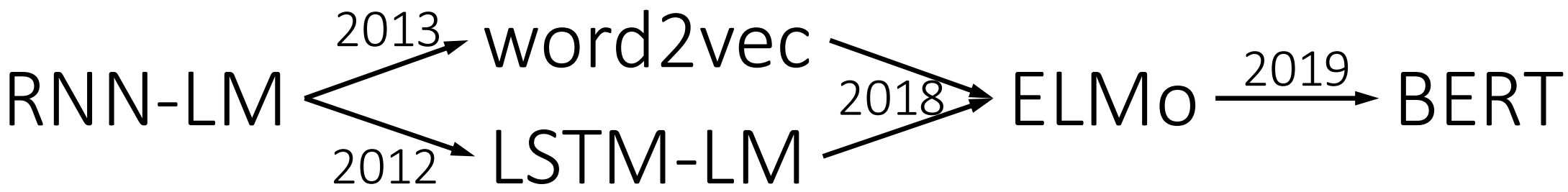
Mikolov et al (2010)



An RNN LSTM Language Model

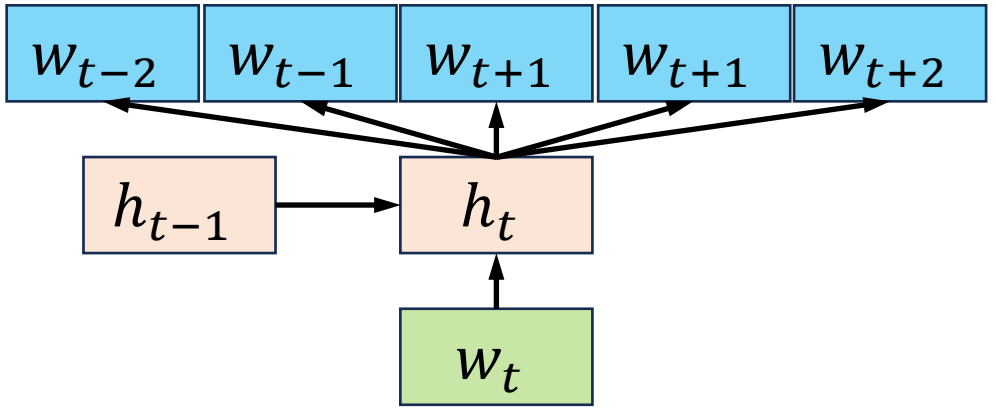
Sundermeyer et al (2012)





word2vec

Distributed Representations of Words and Phrases and their Compositionality



ELMo

Deep contextualized word representations

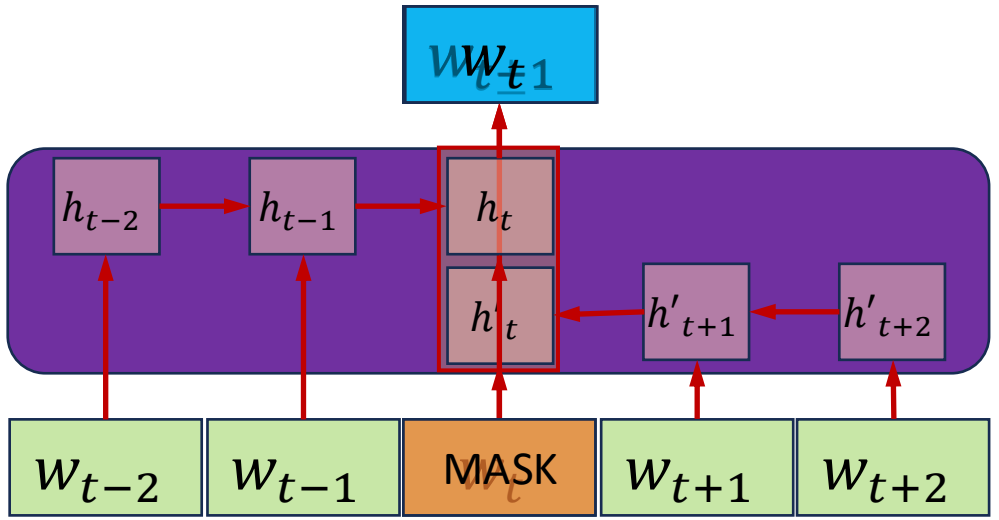
Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

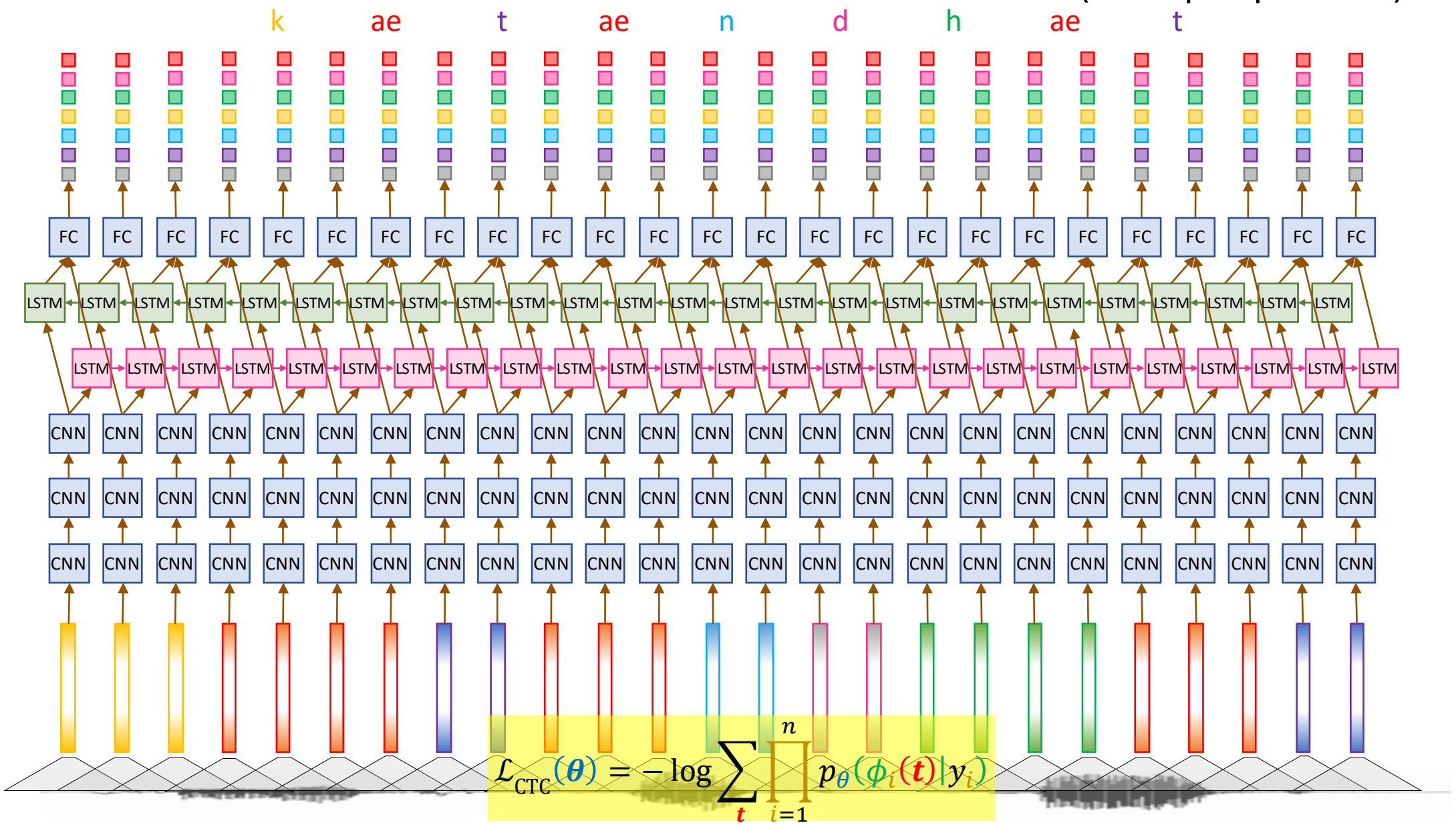
BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

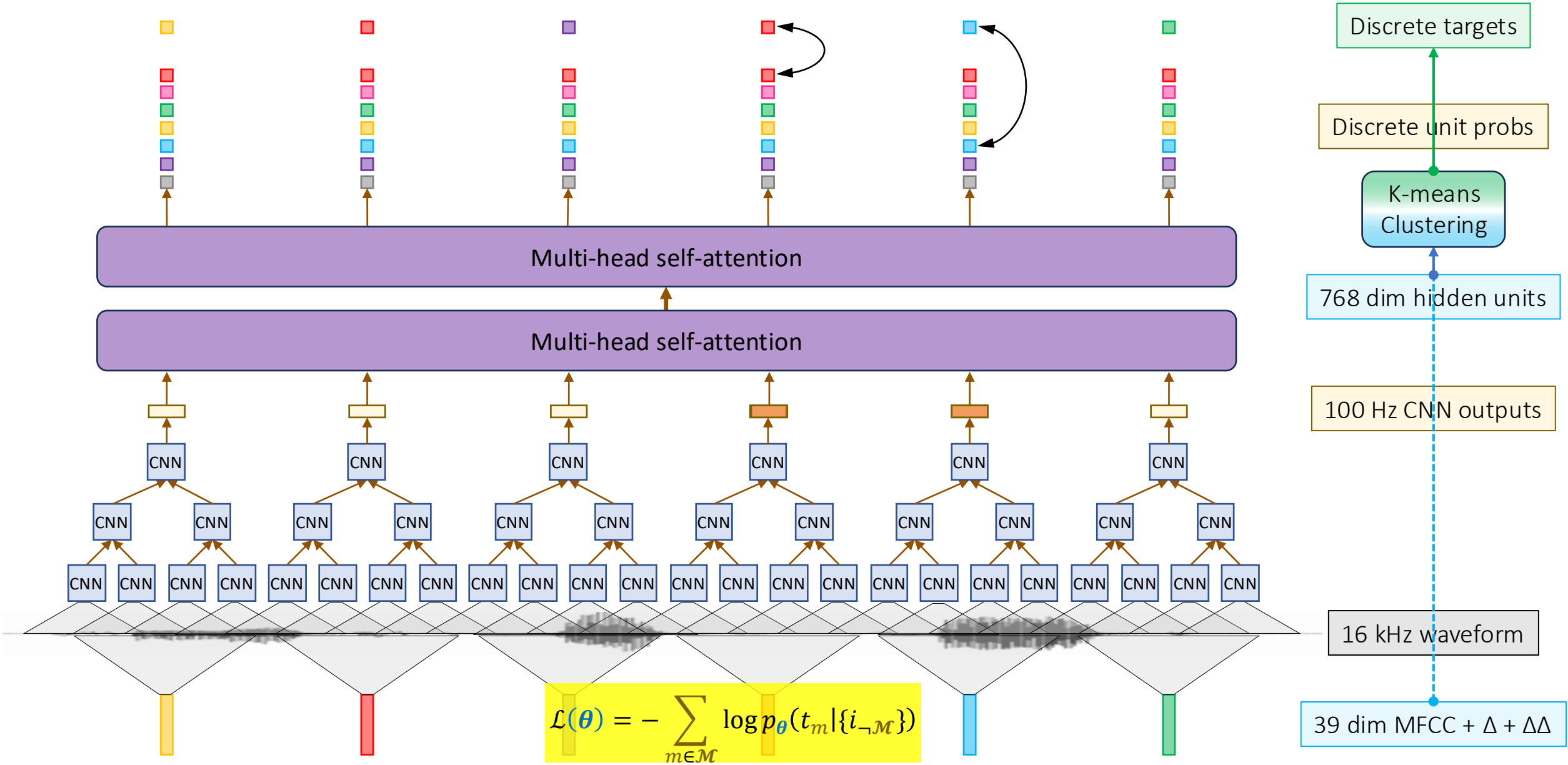
Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com



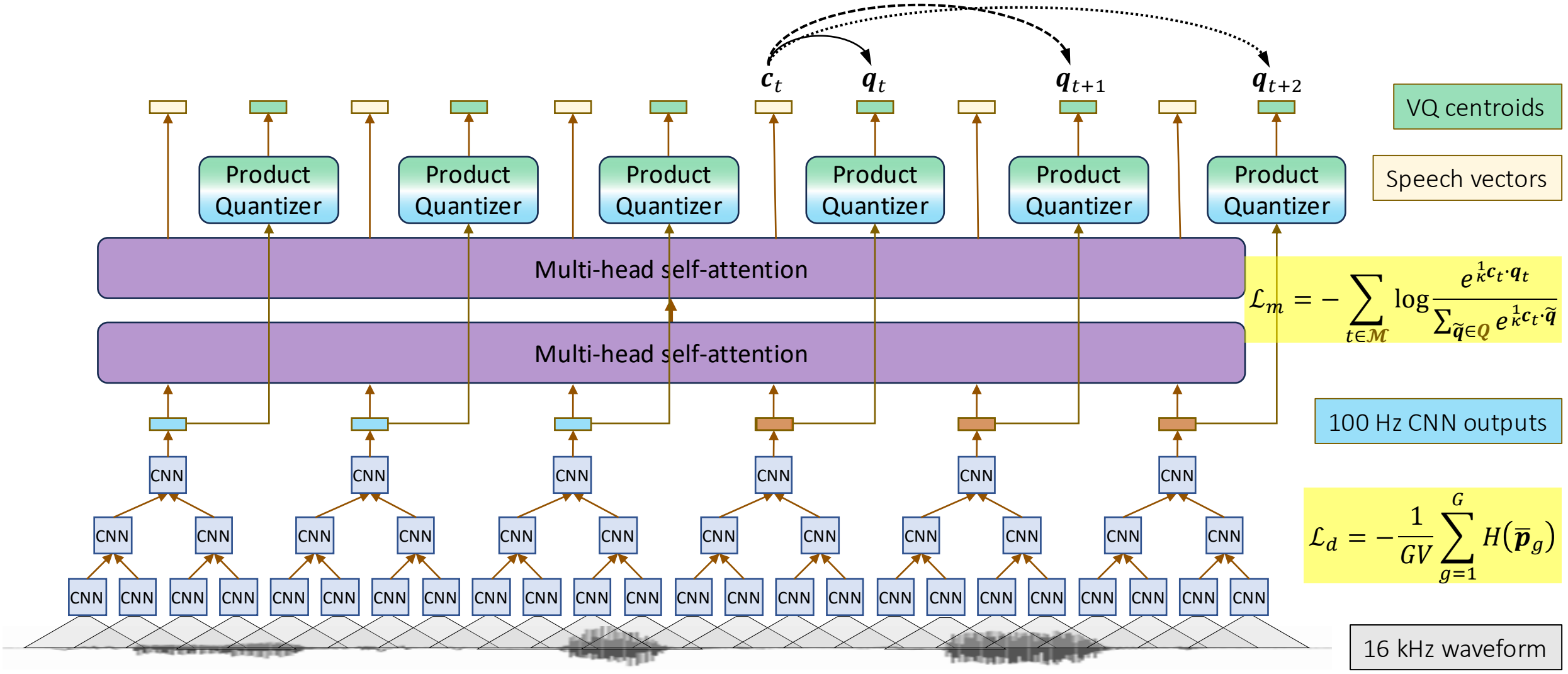
A Bidirectional LSTM Architecture (Deep Speech)



HuBERT – Quantized speech “tokens” and BERT-like loss



wav2vec 2.0 – Noise Contrastive Estimation and Learnt VQ



Fundamentals of Automatic Speech Recognition

Recap

Recap of ASR Fundamentals

- Basic Mathematics
  - Source-Channel Formulation of the ASR Problem
  - Hidden Markov Models
  - Forward-Backward and Viterbi Algorithms
- Neural Methods
  - Neural Acoustic Models (Hybrid ASR Systems)
  - Neural Language Models (RNNs and LSTMs)
  - Connectionist Temporal Classification
  - Encoder-Decoder Models with Attention
  - Recurrent Neural Network Transducers (RNN-T)
- Self-Supervised Representation Learning
  - Masked-token prediction (HuBERT)
  - Noise contrastive estimation (wav2vec 2.0)