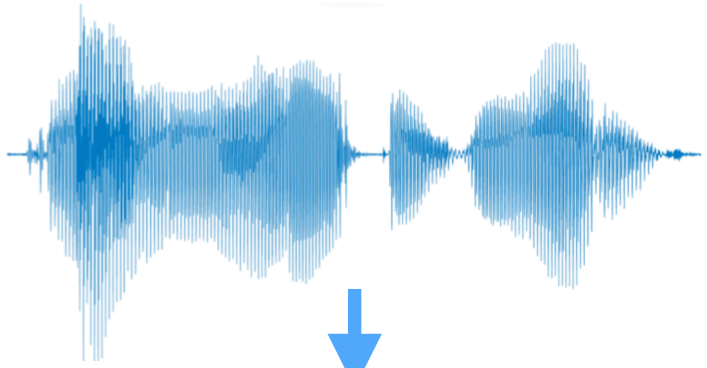
A Fully Differentiable Beam Search Decoder

Ronan Collobert, Awni Hannun, Gabriel Synnaeve

facebook AI Research

Automatic Speech Recognition



acoustic model

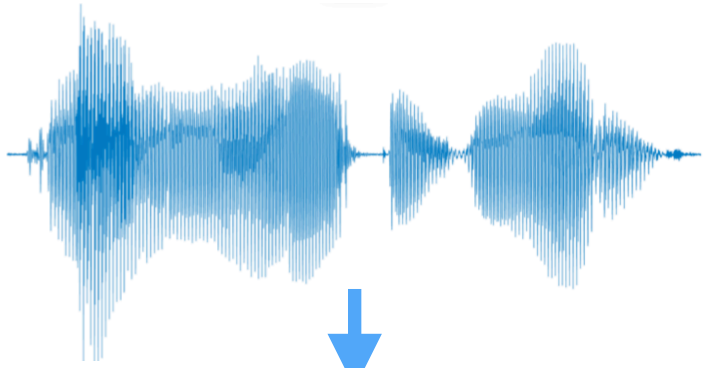
thhhheee ccaaat sssaatttt

Classical Training

Ignore Inference

Acoustic model performs *implicit* language modeling

Automatic Speech Recognition



acoustic model

language model

thhhheee ccaaat sssaatttt

decoder

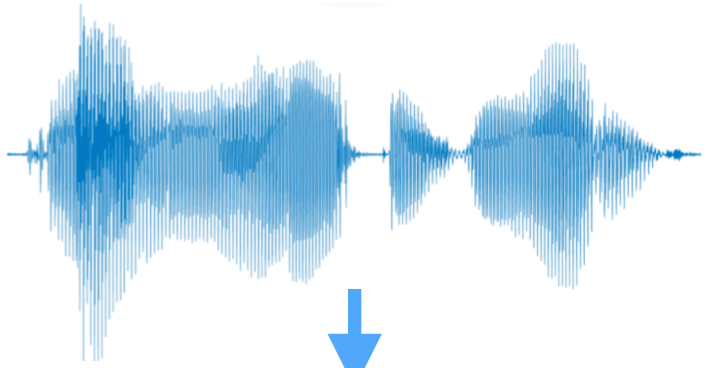
the cat sat

lexicon

Full
End-to-end
Training

Explicit
Language Model

Automatic Speech Recognition



acoustic model

language model

thhhheee ccaaat sssaatttt

decoder

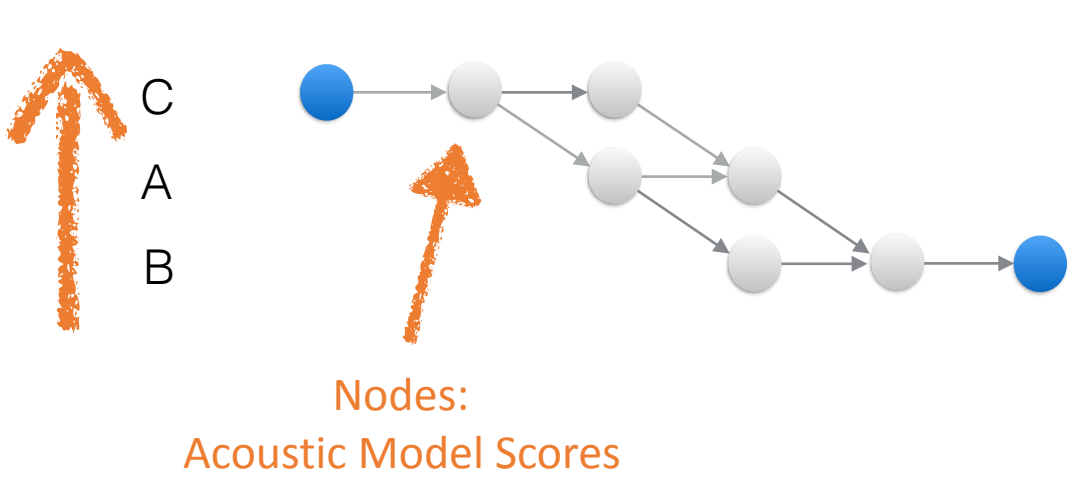
the cat sat

lexicon

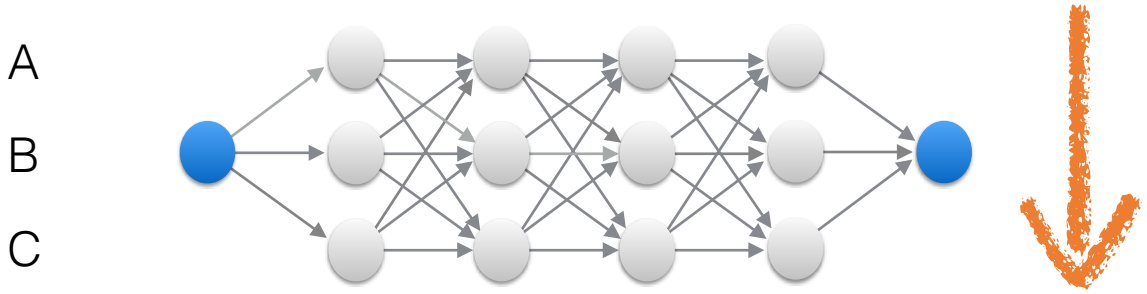
Jointly Train Acoustic and Language Models

Language Model - Free Training

- Say "cab" is the target
  - Dictionary is {a, b, c}
  - Over 4 frames, can be written caab, ccab, cabb, etc..

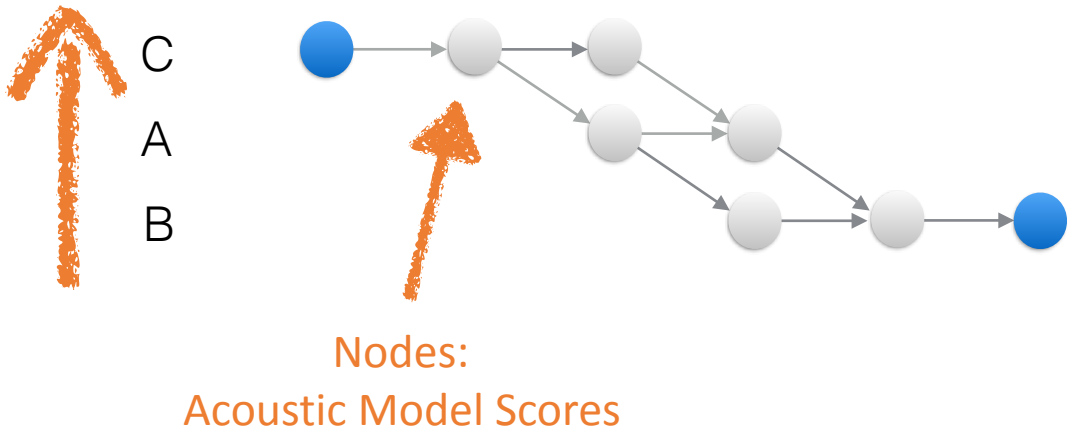


Nodes:
Acoustic Model Scores



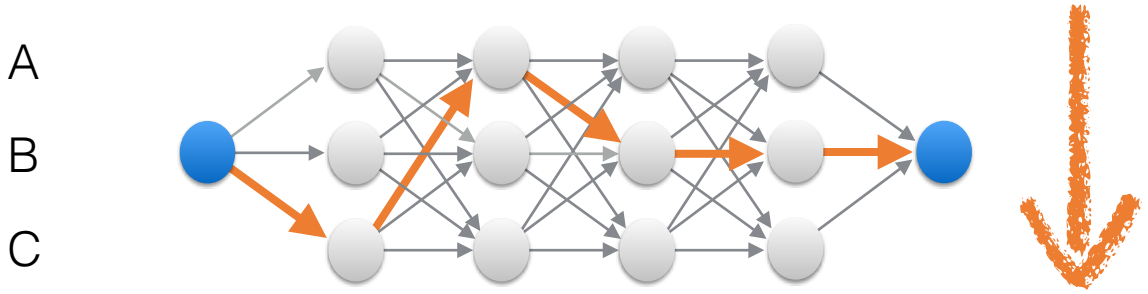
Language Model - Free Training

- Say "cab" is the target
  - Dictionary is {a, b, c}
  - Over 4 frames, can be written caab, ccab, cabb, etc..



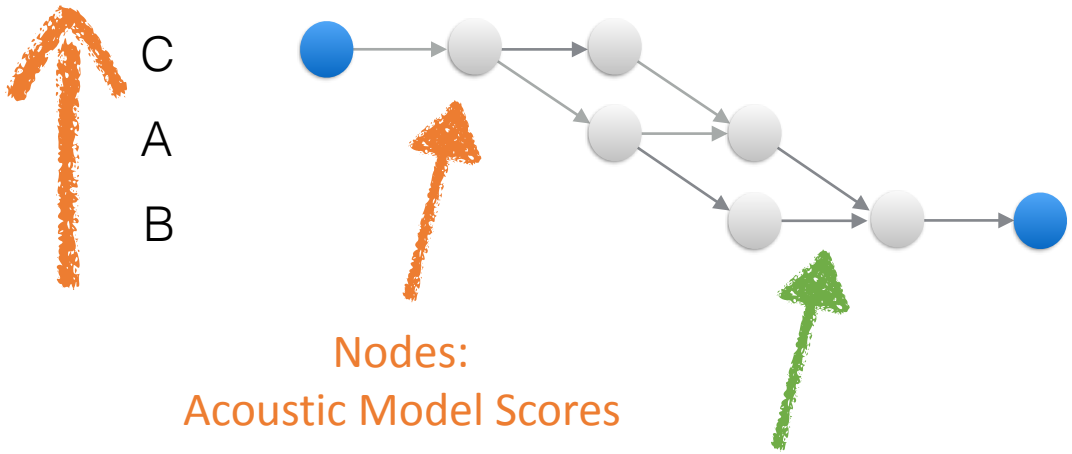
Nodes:
Acoustic Model Scores

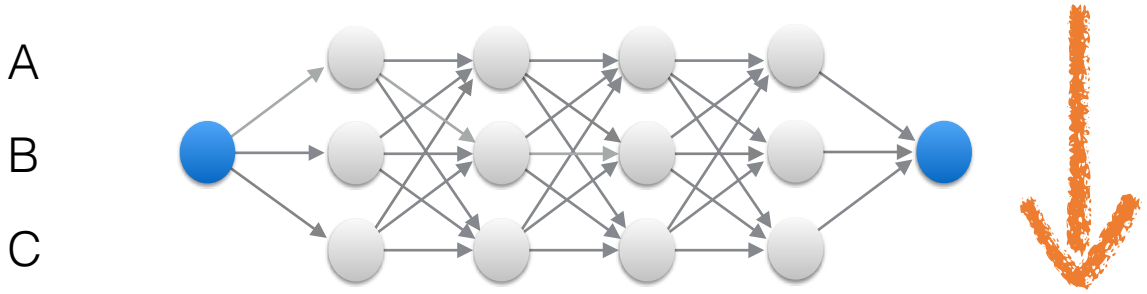
- Viterbi at **inference**



Language Model ~~Free~~ Training

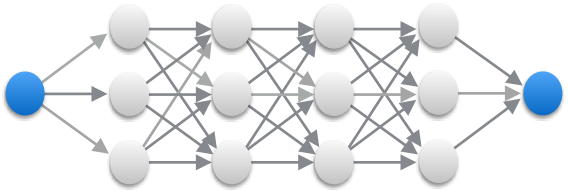
- Say "cab" is the target
  - Dictionary is {a, b, c}
  - Over 4 frames, can be written caab, ccab, cabb, etc..



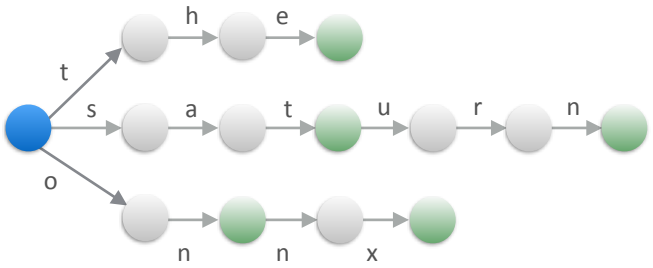


Exact normalization
intractable
for large-context language models

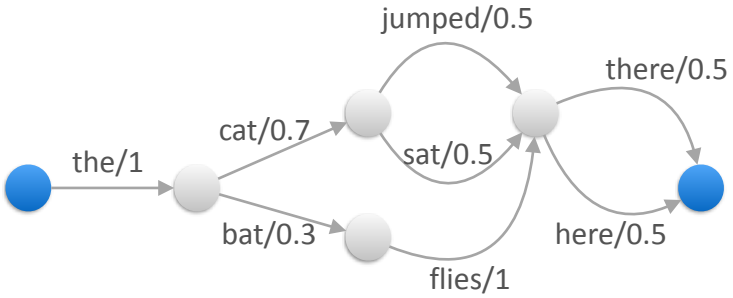
Word-Level Normalization & Decoding



acoustic
A



lexicon
L



word LM
G

- Consider A∘L∘G instead of A
- Beam search! (normalization and inference)

Iterative procedure:

- **Augment** hypothesis constrained to (A, L, G)
- **Merge** *hypothesis* leading to same (L, G) states
- **Prune** *hypothesis*

differentiable!
backprop through
dynamic programming recursion
(but, well, not that simple)

Experiments on WSJ

All results are in Word Error Rate

Pre-trained Language Model

Model	nov93dev	nov92
ASG *10M AM (beam size 8000)*	8.5	5.6
ASG *10M AM (beam size 500)*	8.9	5.7
ASG *7.5M AM (beam size 8000)*	8.8	6.0
ASG *7.5M AM (beam size 500)*	9.4	6.1
DBD *10M AM (beam size 500)*	8.7	5.9
DBD *7.5M AM (beam size 500)*	7.7	5.3
DBD *7.5M AM (beam size 1000)*	7.7	5.1

Grid-search decoding at inference

Trained Language Model (no LM-text data)

Model	nov93dev	nov92
ASG *(zero LM decoding)*	18.3	13.2
ASG *(2-gram LM decoding)*	14.8	11.0
ASG *(4-gram LM decoding)*	14.7	11.3
DBD *zero LM*	16.9	11.6
DBD *2-gram LM*	14.6	10.4
DBD *2-gram-bilinear LM*	14.2	10.0
DBD *4-gram LM*	13.9	9.9
DBD *4-gram-bilinear LM*	14.0	9.8

Smaller beam
Lighter acoustic model
Learn to weight the language model

Can learn non-trivial LM

A Fully Differentiable Beam Search Decoder
Ronan Collobert, Awni Hannun, Gabriel Synnaeve