Google

Research & Machine Intelligence

Natural Language Processing with Small Feed-Forward Networks

Jan A. Botha **Emily Pitler** Ji Ma Anton Bakalov
Alex Salcianu David Weiss Ryan McDonald Slav Petrov

emnlp 2017
Copenhagen

Goal: Small and Fast (on CPU/mobile)

Deep recurrent models: can have 100s of millions (or billions) of parameters

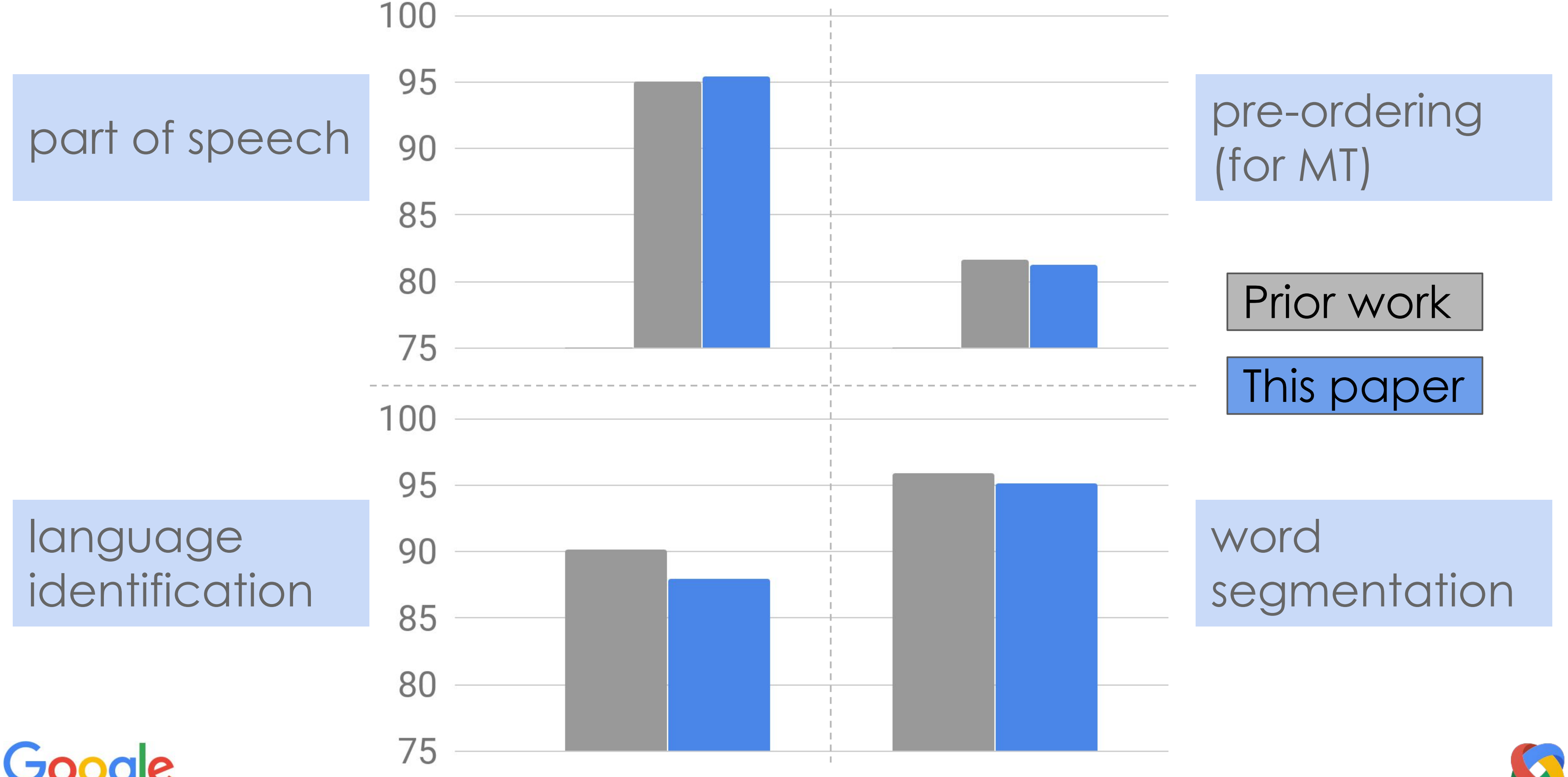
Recent work on smaller recurrent models: Kim and Rush, 2016; Sharp Models on Dull Hardware, Devlin, EMNLP 2017 10's to 100 tokens/second

For variety of NLP tasks, can get order of magnitude speedup over LSTMs

Memory	**≤ 2 MB**
Speed	**7k – 46k** tokens/second
Trained	**in a few hours**

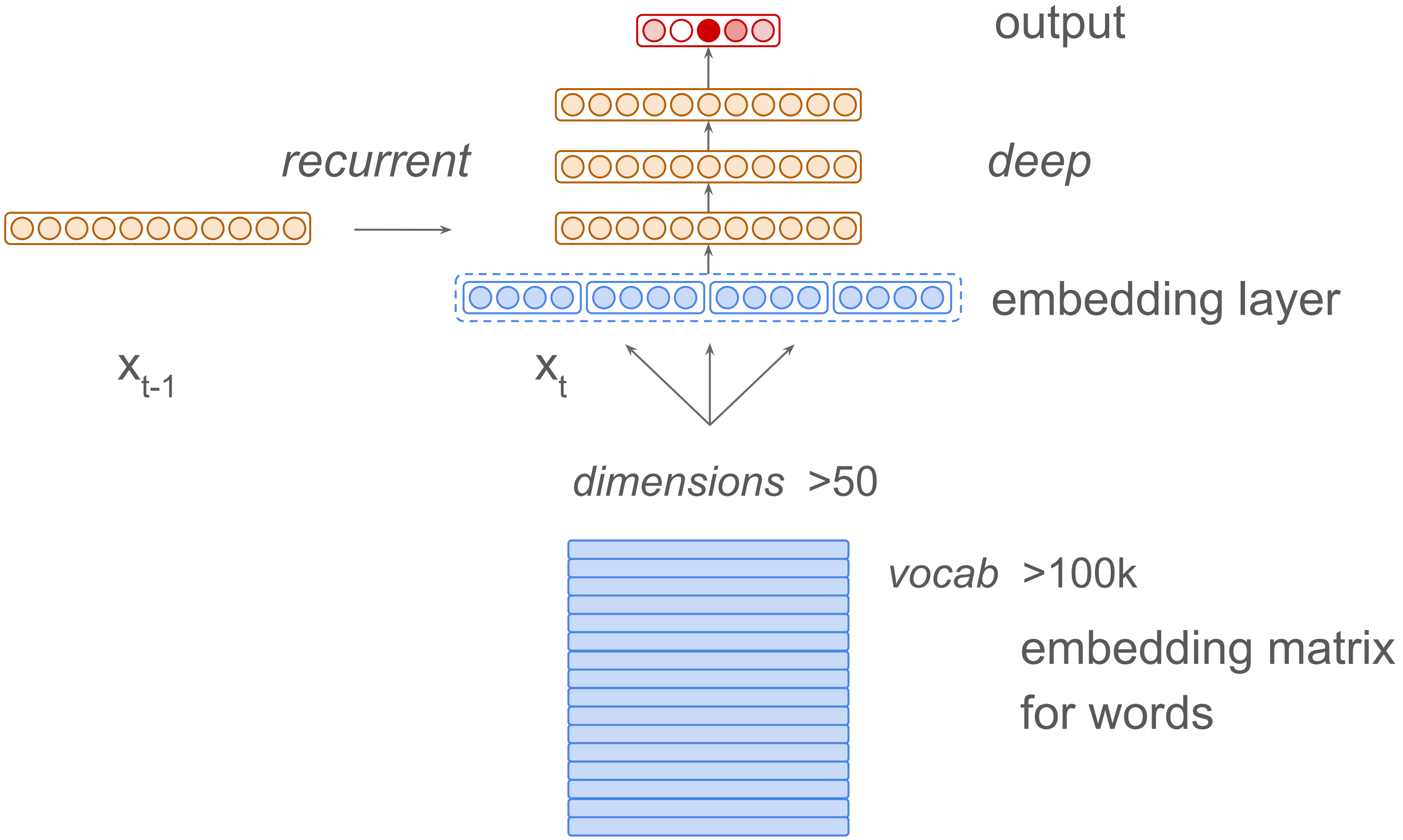
Google

...With Near State-of-the-Art Accuracies in 4 Tasks

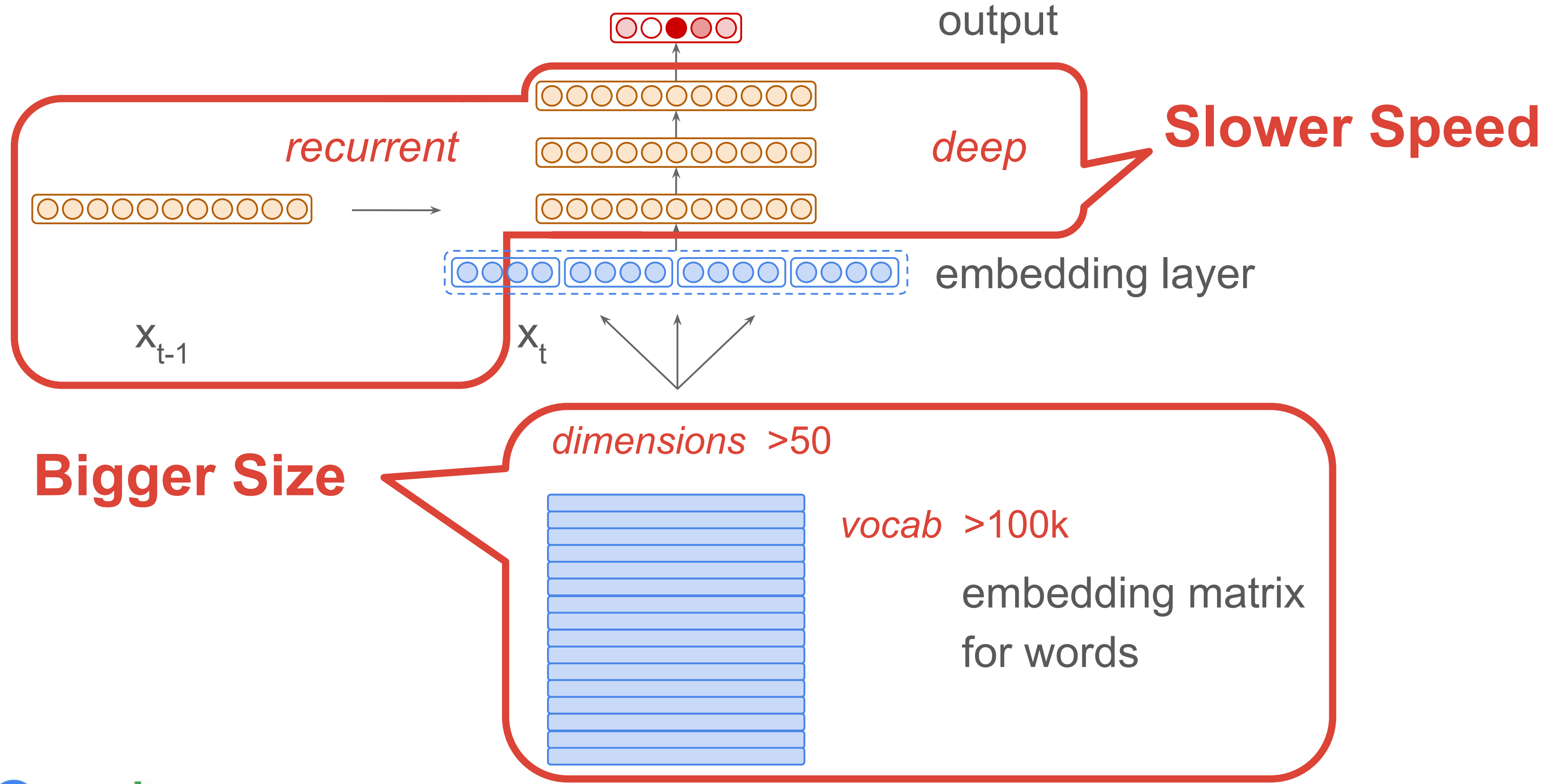


Google

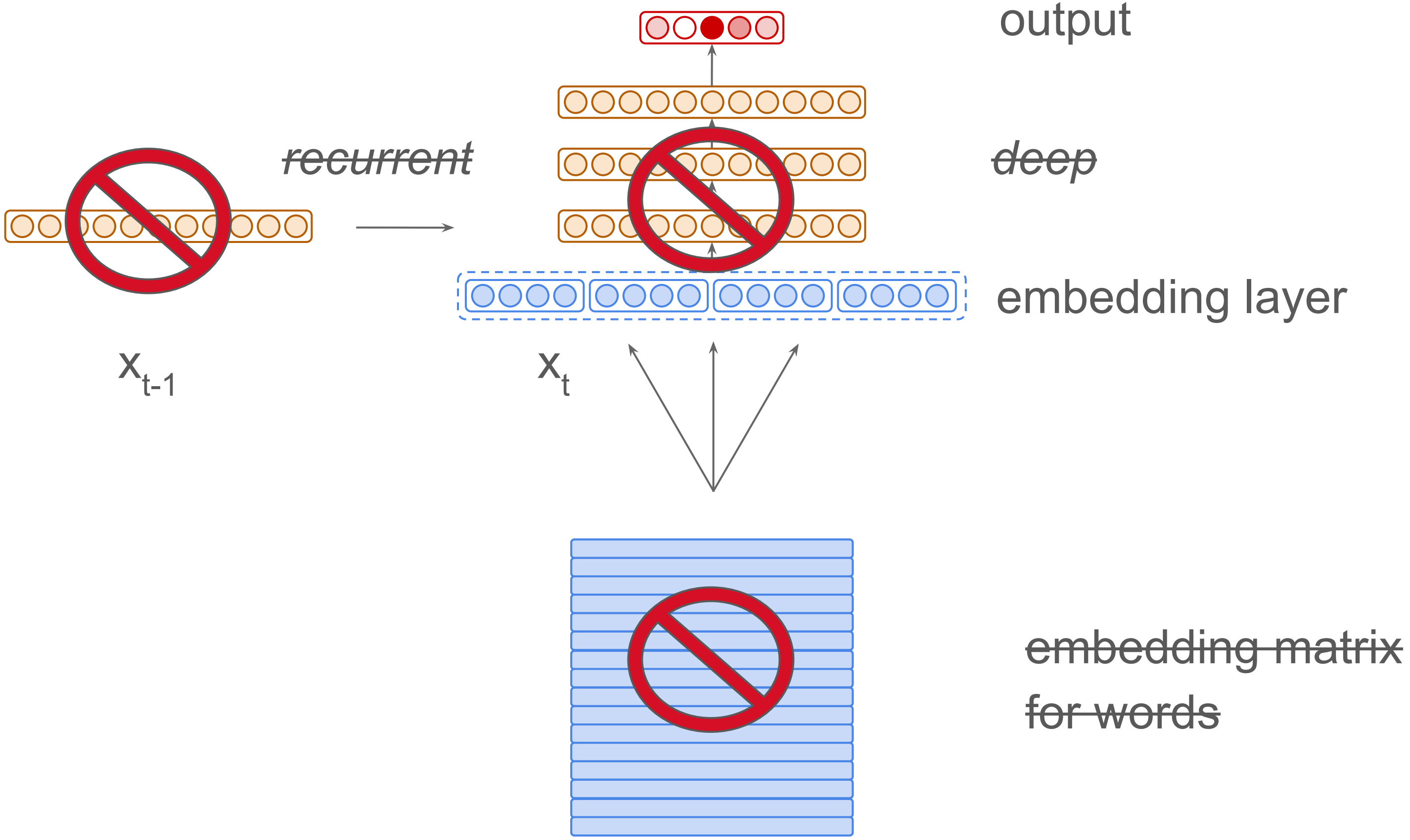
A Recurrent & Deep Model with Large Vocabulary



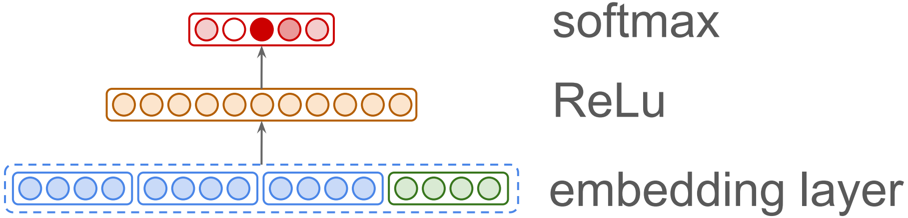
A Recurrent & Deep Model with Large Vocabulary



How Best to Allocate a Small Memory Budget?



What’s Left? Small Feed-Forward Architecture



Goals

Accuracy

Model Size

Word Clusters
Pipelines
Selected Features

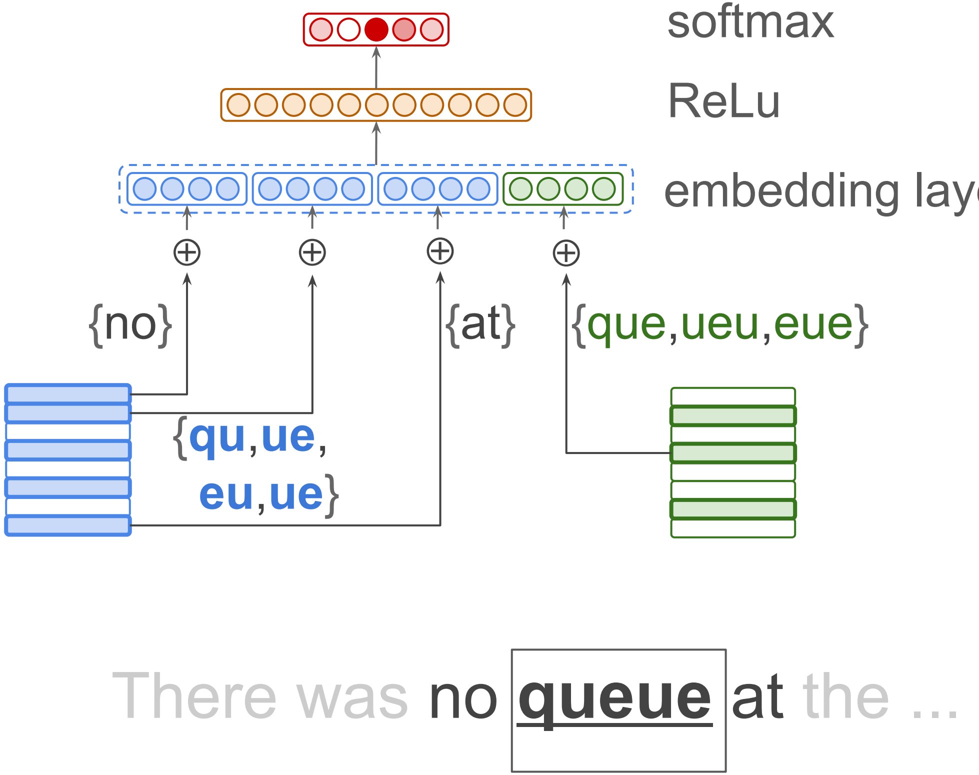
Hashed Character n-grams
Quantization

Input Representation: Hashed Character n-grams

Extract ***char-ngrams*** *at different offsets,*

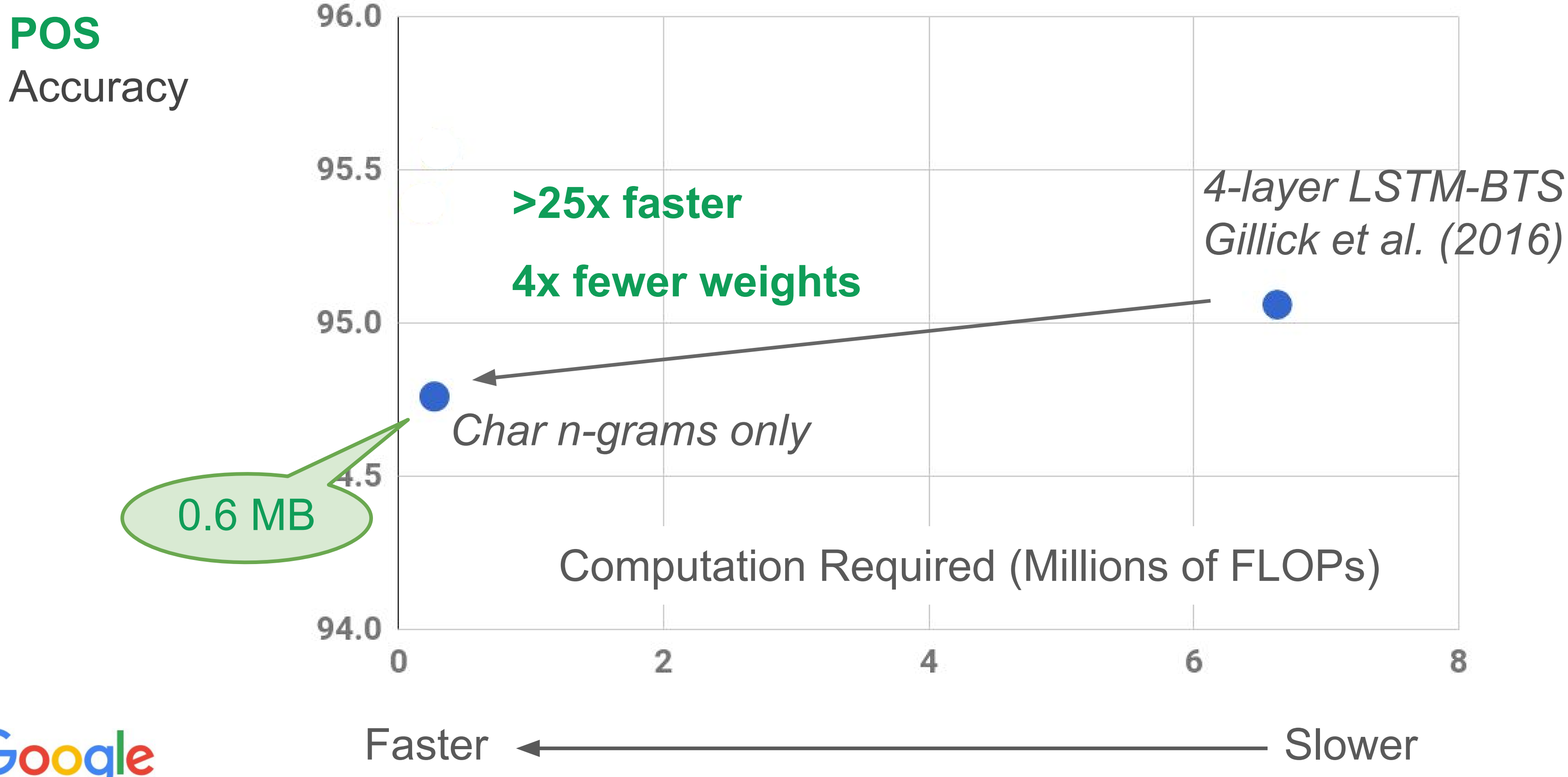
apply ***hash****,*

aggregate *looked up* ***embeddings.***

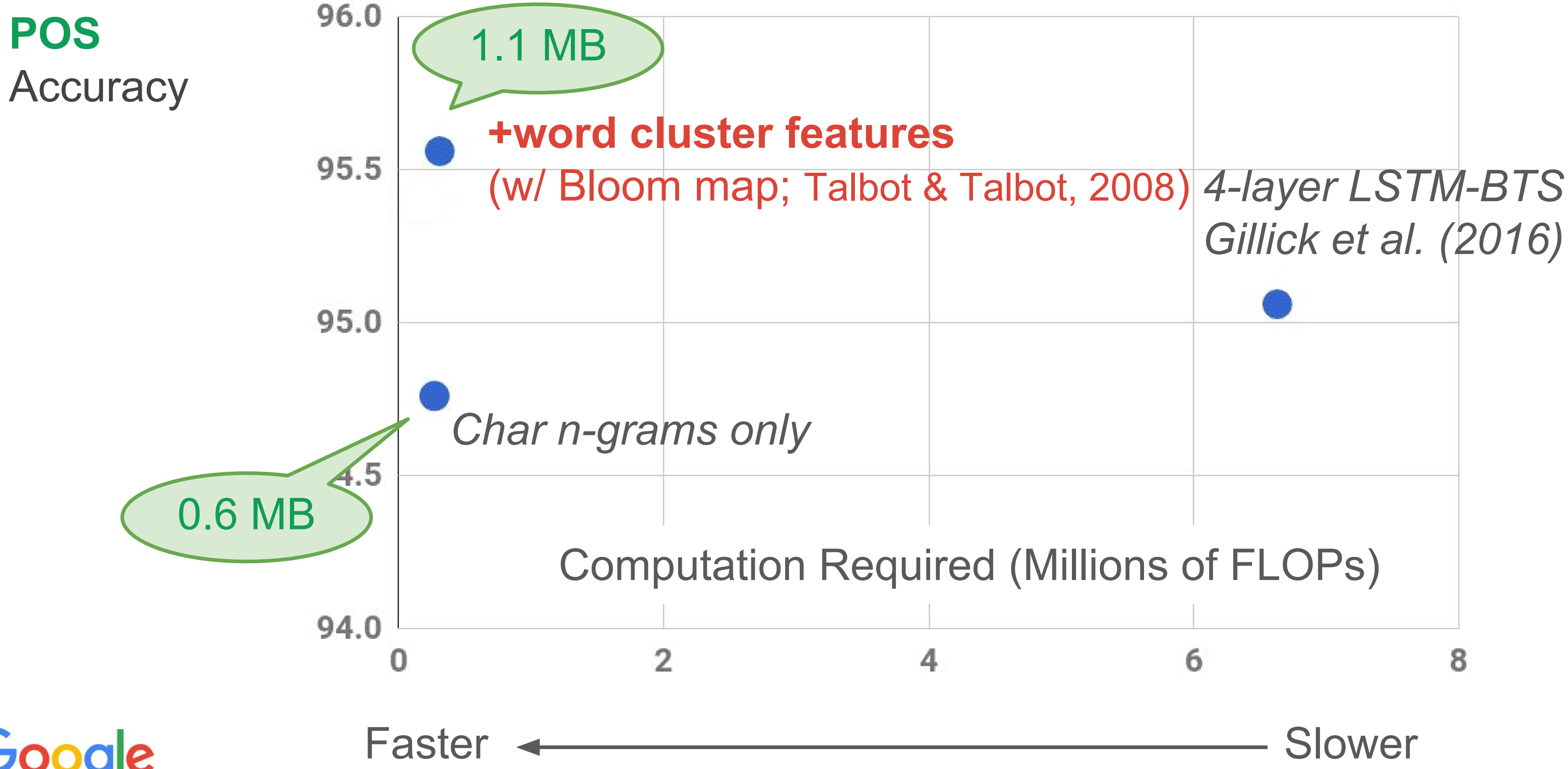


Case Study 1: POS Tagging

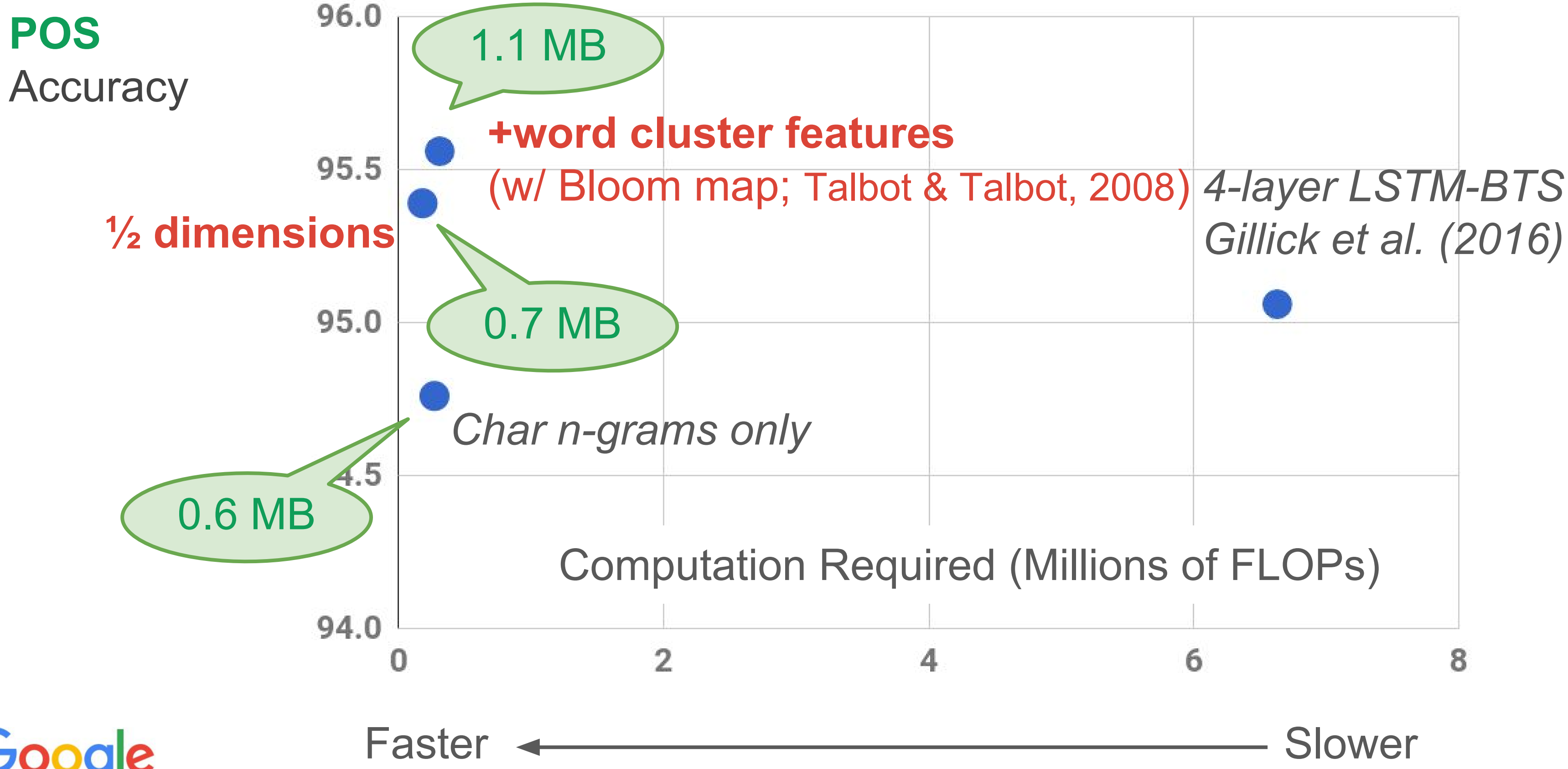
Vanilla Model: Less Resources, Little Less Accurate



Accuracy Boost Adding Resource-backed Features



Allows Reducing Embedding Dimensions Further



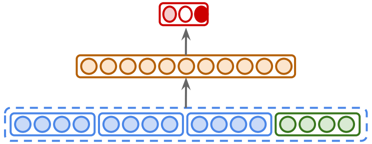
Case Study 2: Preordering for MT

Transition System for Structured Output

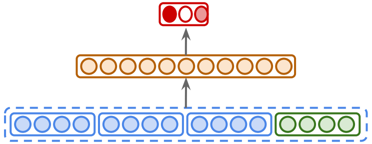
English → Japanese word ordering: I ate pizza → I pizza ate

Stack	Buffer
I ate pizza	
I pizza	ate

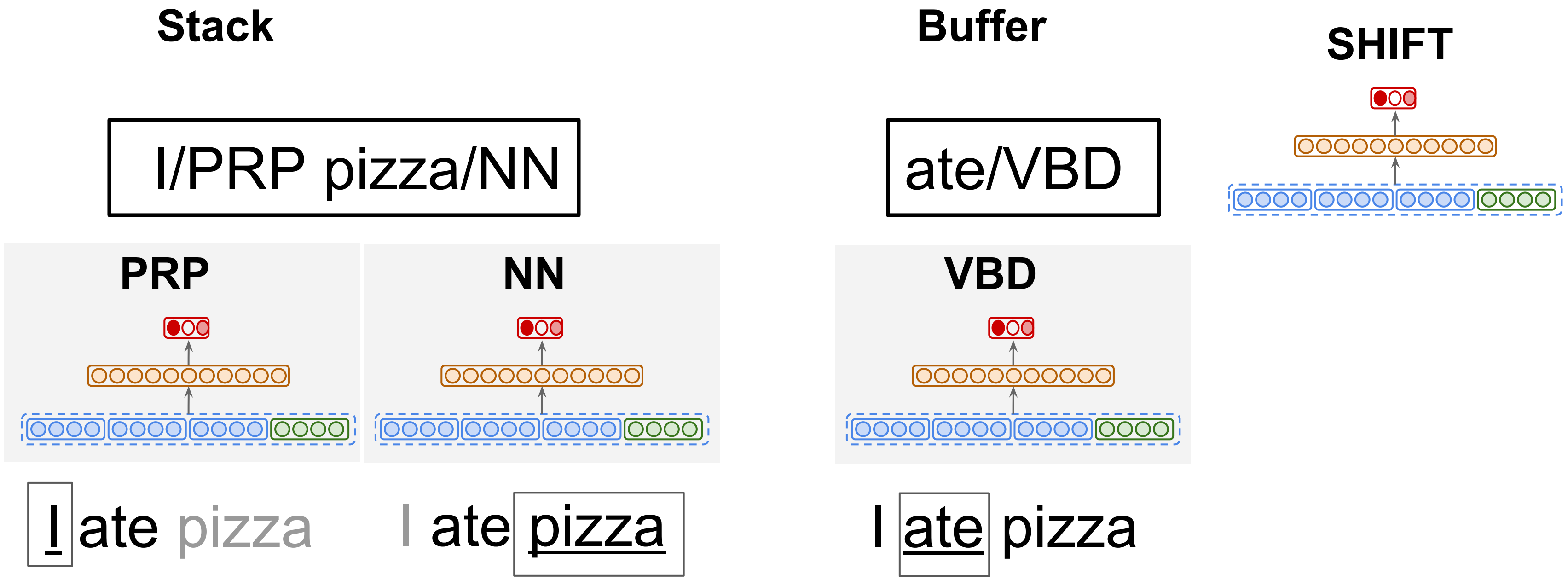
SWAP



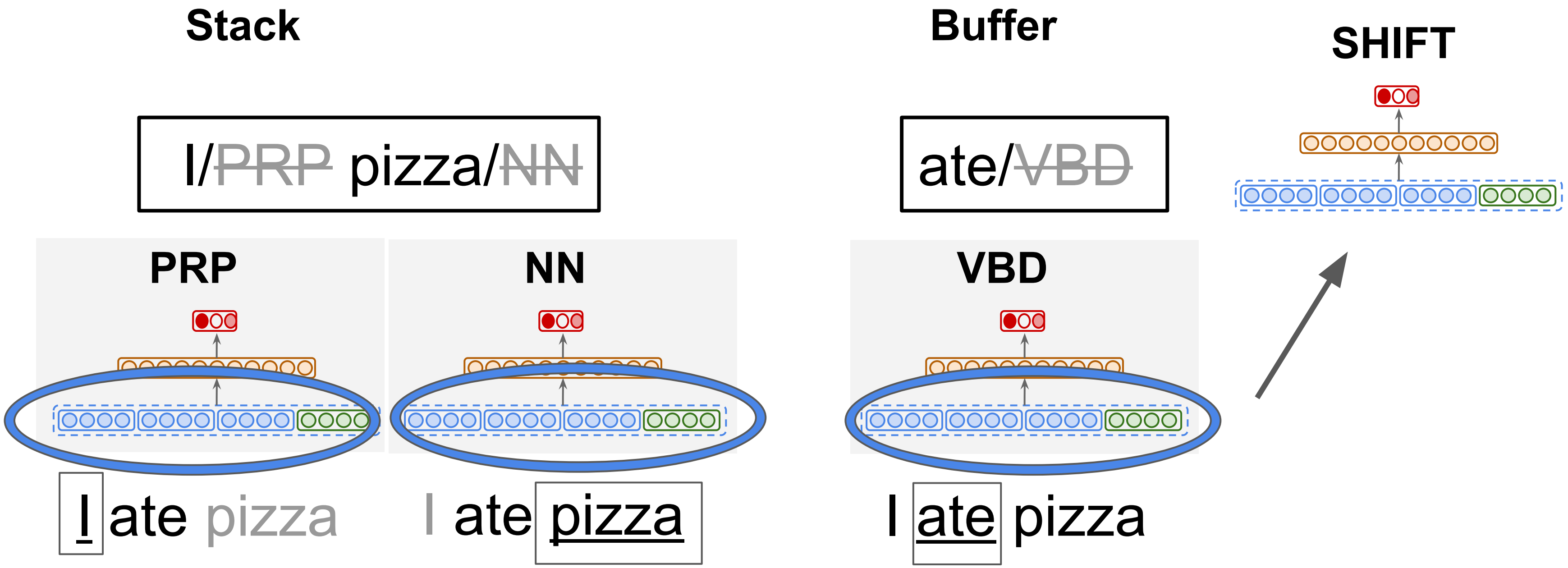
SHIFT



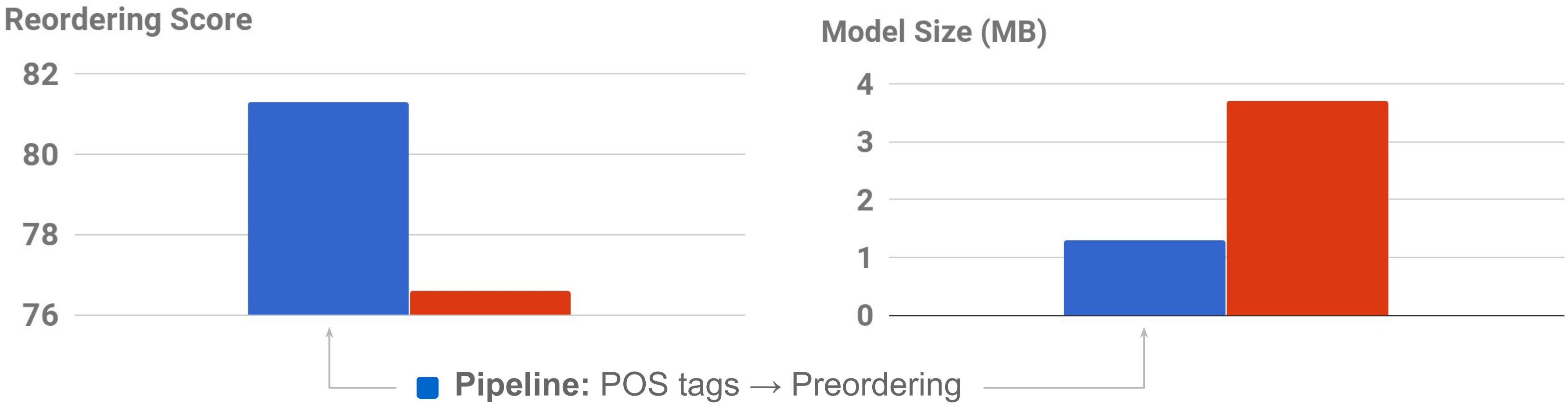
POS Tags as Intermediate Representation



Or, “End-to-End Style” with Features from Tagger



Pipeline: Higher Accuracy for Smaller Size



Pipeline: POS tags → Preordering

End-to-end: Preordering+Tagger features

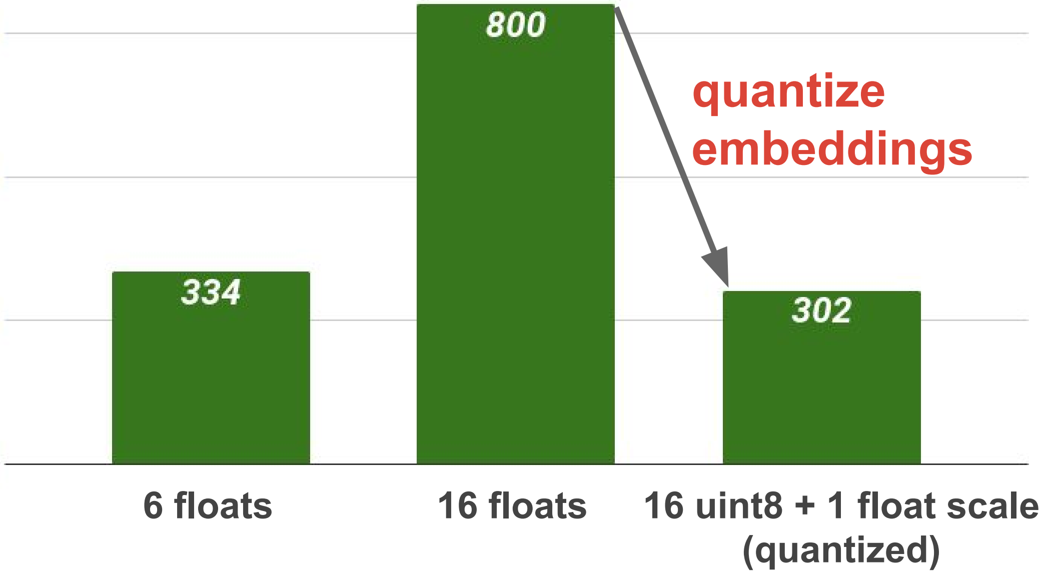
Case Study 3: Language Identification

LangID: Post-hoc Quantization to Reduce Space

Baldwin & Lui (2010)
.870 - .902

F1: .873 .880 .880

Model size (KB)



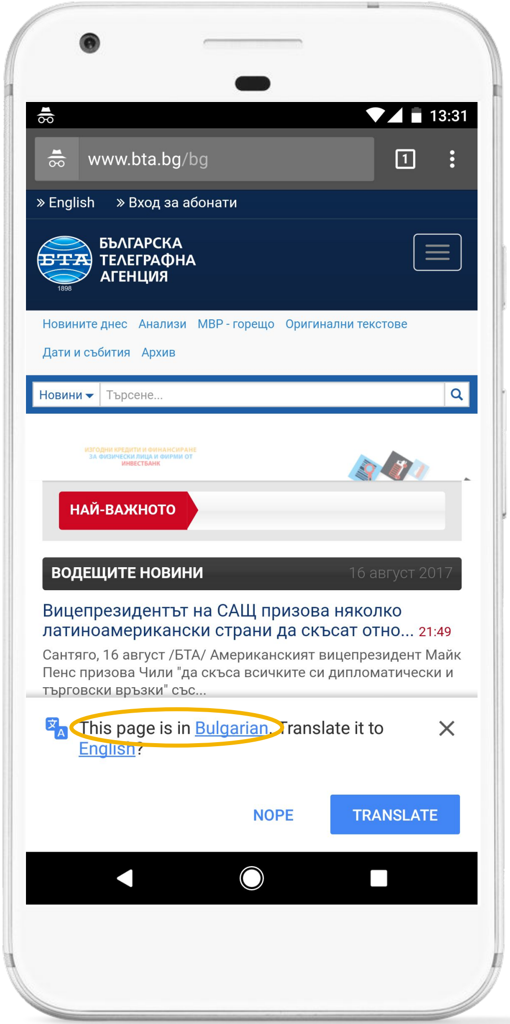
6 dimensions **16 dimensions**

Google

That LangID model is essentially...

Compact Language Detector v3 (CLD3)

- ✓ runs inside all Google Chrome browsers
- ✓ code: github.com/google/CLD3



Actual screenshot from 16 Aug 2017

Conclusion

Small (<= 2 MB) & fast (7k - 46k tokens/second) models with high accuracy in multiple tasks

Explicit intermediate representations & engineered features bring big accuracy gains in low-memory setting

Simple techniques → easy to include in standard practice