Modeling Non-Linguistic Contextual Signals in LSTM Language Models via Domain Adaptation

Min Ma* Shankar Kumar† Fadi Biadsy† Michael Nirschl† Tomas Vykruta† Pedro Moreno†
* Graduate Center, The City University of New York † Google Inc., New York, NY, USA

Introduction

Language Models (LMs) for Automatic Speech Recognition (ASR) can benefit from utilizing non-linguistic contextual signals such as application (app) ids

The vast majority of speech queries lack annotations of such signals, making it challenging to directly train domain-specific LMs

We propose three domain adaptation schemes to improve the domain-level performance of Long Short-Term Memory (LSTM) LMs in pre-training & fine-tuning stages

Data & Baselines

Three utterance-level app signals:

- Google Maps (27.6%)
- Google PlayStore (24.4%)
- YouTube (48.0%)

Sets	#(utts))	#(words)
OOD-Train	257M	1.6B
OOD-Dev	53K	343K
DOM-Train	23M	109M
DOM-Dev	6.8M	31M

Distribution of out-of-domain/domain data

Baselines: 2-layer * 1024-node LSTM LMs, each word is embedded in 1024 dimensions

- One-pass baselines: Train the LMs on OOD-Train only or DOM-Train only
- Fine-tuning baseline: Pre-train on OOD-Train, fine-tune models on DOM-Train

Baselines	DOM	OOD
OOD-Data Only Baseline	82	85
DOM-Data Only Baseline	49	293
Fine-tuning Baseline	47	209

Perplexity of baseline models

Scheme I: Prepend App Id

- Initialize LSTM state by contextual app id
- Learn app-embeddings by back-propagation
- **17.0%** rel. reduction in domain perplexity

Adaptation Strategy	DOM	OOD
Fine-tuning Baseline	47	209
Adaptation Scheme I	**39**	211
Sch I, freeze Embed [1]	41	176
Sch I, freeze Embed, LSTM layers	46	153

Scheme II: Add Meta-Memory

- Add new layer to embed the app id
- META-MEMORY: Apply affine transformations to the app embedding **a**
- Add META-MEMORY to LSTM cells (omitting the biases)

$$
\begin{aligned}
f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + \mathbf{W_{fa}}\mathbf{a}) \\
i_t &= 1 - f_t \\
\hat{c}_t &= tanh(W_{cx}x_t + W_{ch}h_{t-1} + \mathbf{W_{ca}}\mathbf{a}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \hat{c}_t \\
o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t + \mathbf{W_{oa}}\mathbf{a}) \\
h_t &= o_t \odot tanh(c_t)
\end{aligned}
$$

Freezing variants: freeze word embedding layer (and the original LSTM parameters LSTMs*)

Cand variant: only include META-MEMORY in the computation of cell state candidate $\hat{c}_t$

Adaptation Strategy	DOM	OOD
Background LM	83	84
Adaptation Scheme II	**37**	602
Sch II, freeze Embed	41	176
Sch II, freeze Embed, LSTMs*	44	139
Cand Variant of Sch II	**39**	276
Cand, freeze Embed	48	177
Cand, freeze Embed, LSTMs*	51	123

Scheme III: Learn Domain-Parameters in Adaptation Phase

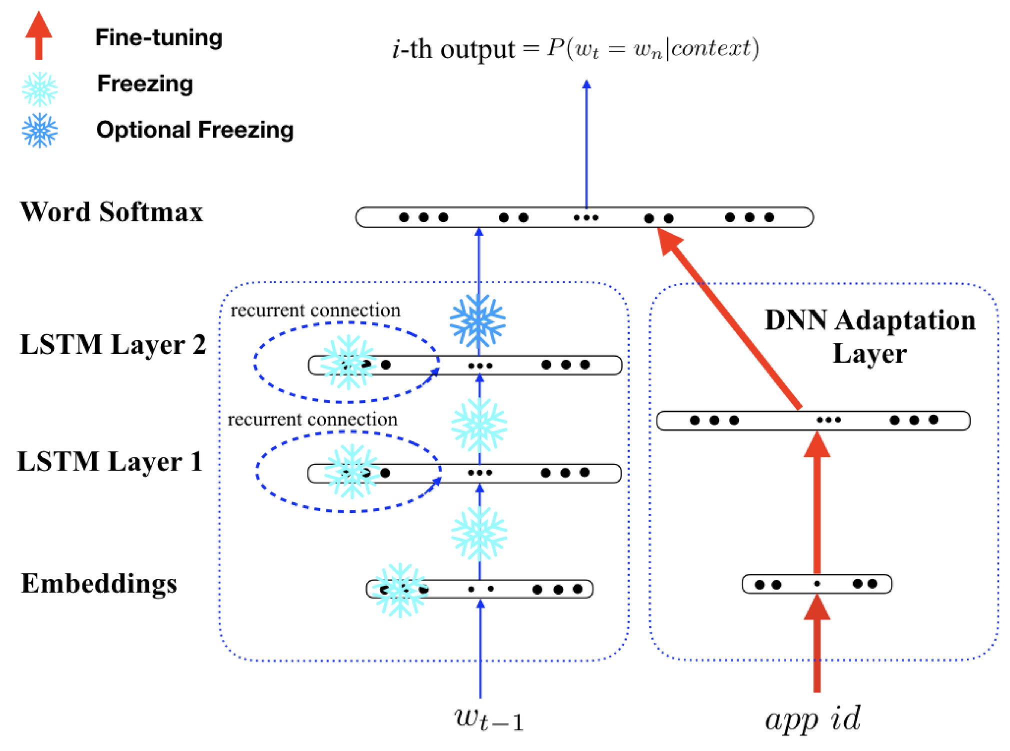
Categorize the model parameters into two sets [2]:

- **Pre-training**: tune general (non-domain) parameters: Word Embed, LSTMs, and W_{OOD}, b_{OOD}
- **Adaptation**: freeze general parameters and tune domain parameters: appEmb, DNNadapt, and W_D, b_D

$$
P(w_t|hist) = \begin{cases} \phi(W_{OOD}h_t + b_{OOD}) \\ \phi(W_{OOD}h_t + b_{OOD} + W_D {h_t}' + b_D) \end{cases}
$$

Variants of Scheme III: relax freezing constraints to fine-tune W_{OOD} and b_{OOD}, but multiply their gradients by a factor

Adaptation Strategy	DOM	OOD
Background LM (OOD Baseline)	82	**85**
Adaptation Scheme III	75	**85**
Variant 1 (mul 0.25)	62	132
Variant 2 (mul 0.50)	61	152
Variant 3 (mul 0.75)	60	152
Variant 4 (mul 1.00)	**51**	133



Learning domain parameters in the adaptation

ASR Results & Conclusions

Language Model	General	Domain
OOD-Data Only Baseline	12.9	13.2
DOM-Data Only Baseline	13.4	13.2
Fine-tuning Baseline	12.9	12.9
Scheme I (prepend)	**12.9**	**12.8**
Scheme I mixed data	**12.9**	**12.8**
Scheme II (meta-memory)	13.0	12.9
Scheme III (dual paths)	13.0	12.9
Scheme III, no penalty	12.9	13.0

WER results

Task	Win/Loss	Change	p-value
PlayStore	58/36	2.4	**0.1%-0.5%**
Maps	82/49	0.9	**1.0%-2.0%**
YouTube	38/27	2.4	5.0%-10.0%

SxS results for three app domains (Scheme I)

- Adding contextual signals to LSTM LM reduces domain perplexity by **21%** relative
- 3% relative reduction in WER on top of an unadapted 5-gram LM
- SxS experiments show significant improvements on sub-domains
- Grouping model parameters into two sets suggests a possible solution to catastrophic forgetting

References:

[1] Ma et al. Approaches for neural-network language model adaptation. In *INTERSPEECH*, 2017.

[2] Biadsy et al. Effectively building tera scale maxent language models incorporating non-linguistic signals. In *INTERSPEECH*, 2017.