Gaussian Process LSTM Recurrent Neural Network Language Models for Speech Recognition

Max W. Y. Lam* Xie Chen† Shoukang Hu* Jianwei Yu* Xunying Liu* Helen Meng*

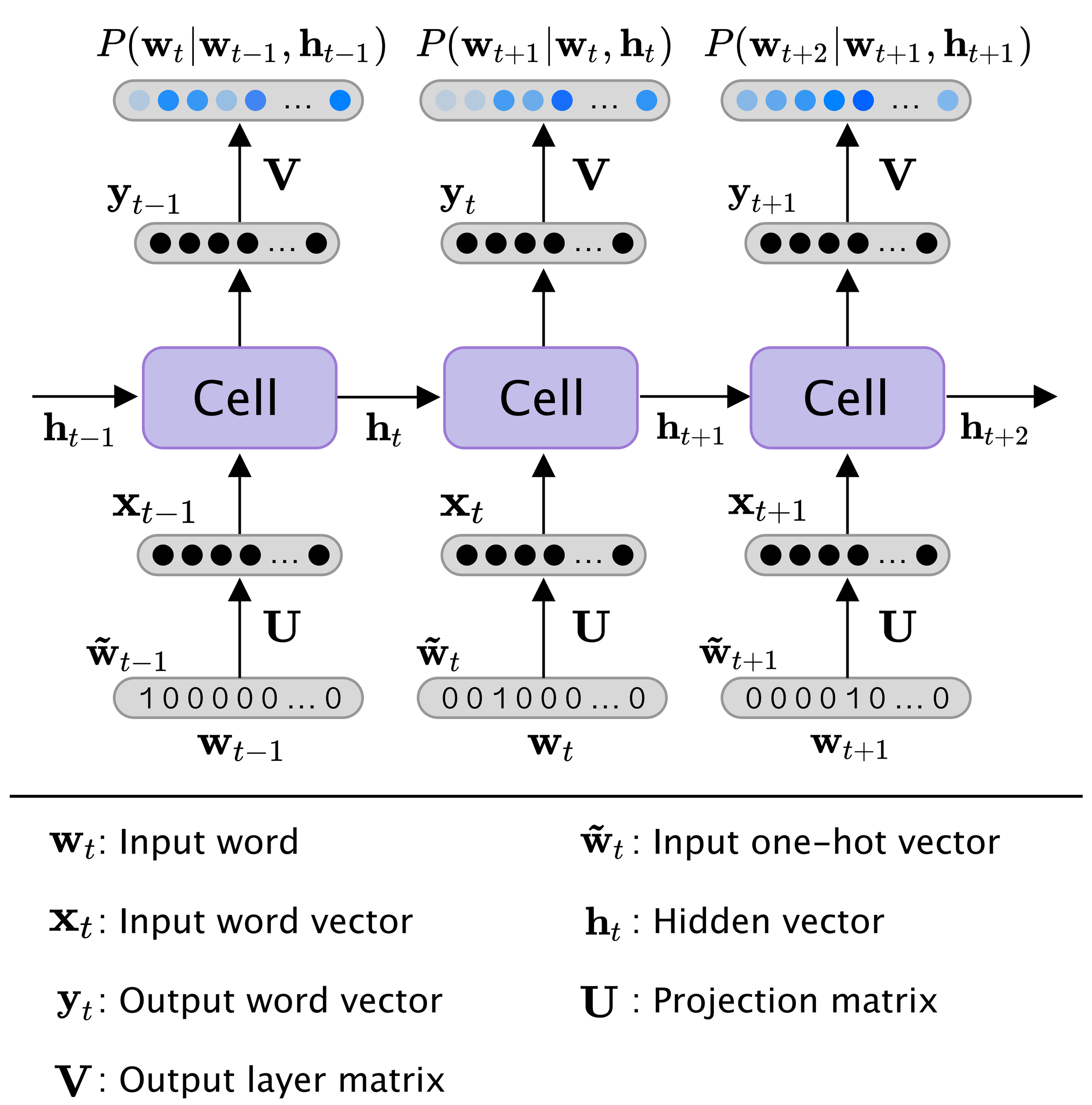
*The Chinese University of Hong Kong, Hong Kong SAR, China

†Microsoft AI and Research, One Microsoft Way, Redmond, WA, USA

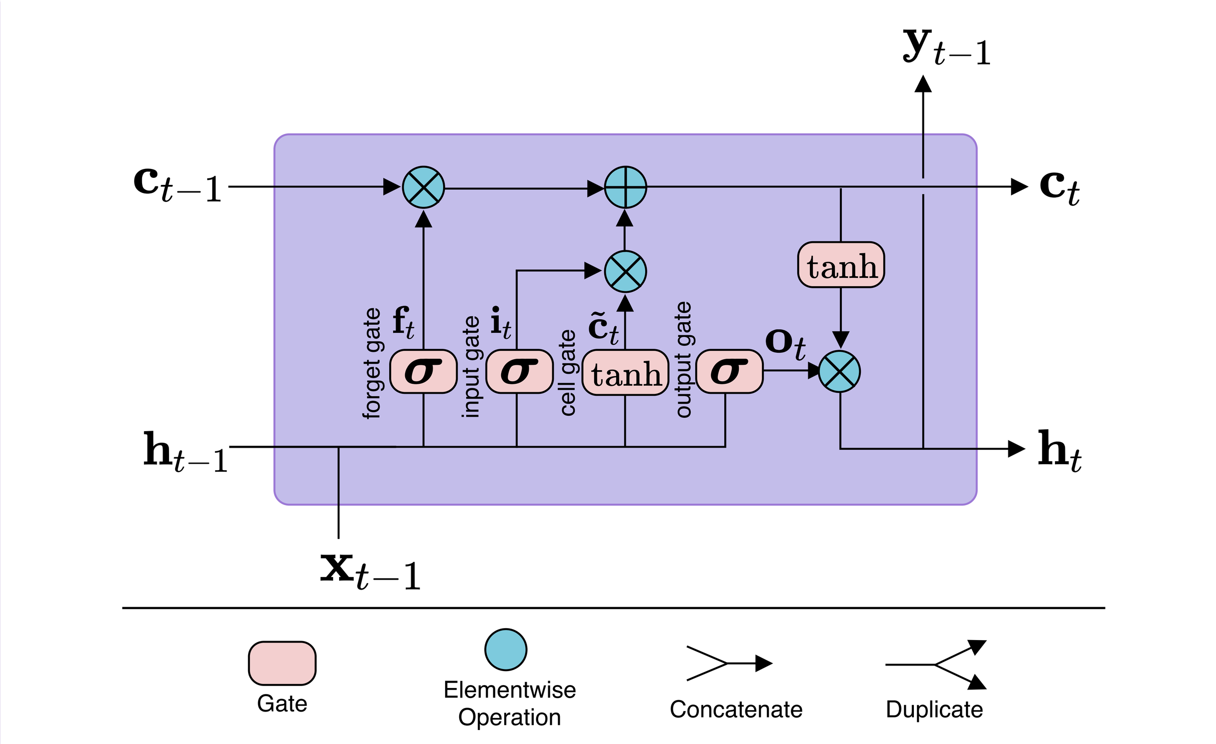
INTRODUCTION

- **Objective:**
  Improve the state-of-the-art LSTM Recurrent Neural Networks (RNNLMs) in ASR
- **Standard LSTM RNNLMs:**
  1) The same form of activation functions for all nodes in each cell
  2) Deterministic weight parameter estimates
- **Limitations:**
  1) Need flexibly optimized activation functions for memory gating given different datasets
  2) Prone to over-fitting and poor generalization on limited training data
- **Proposed GP-LSTM RNNLMs:**
  1) Adopt Gaussian process (GP) to model the uncertainty of acitvation functions
  2) Automatically learn the optimal forms of gates for all hidden nodes in each LSTM cell

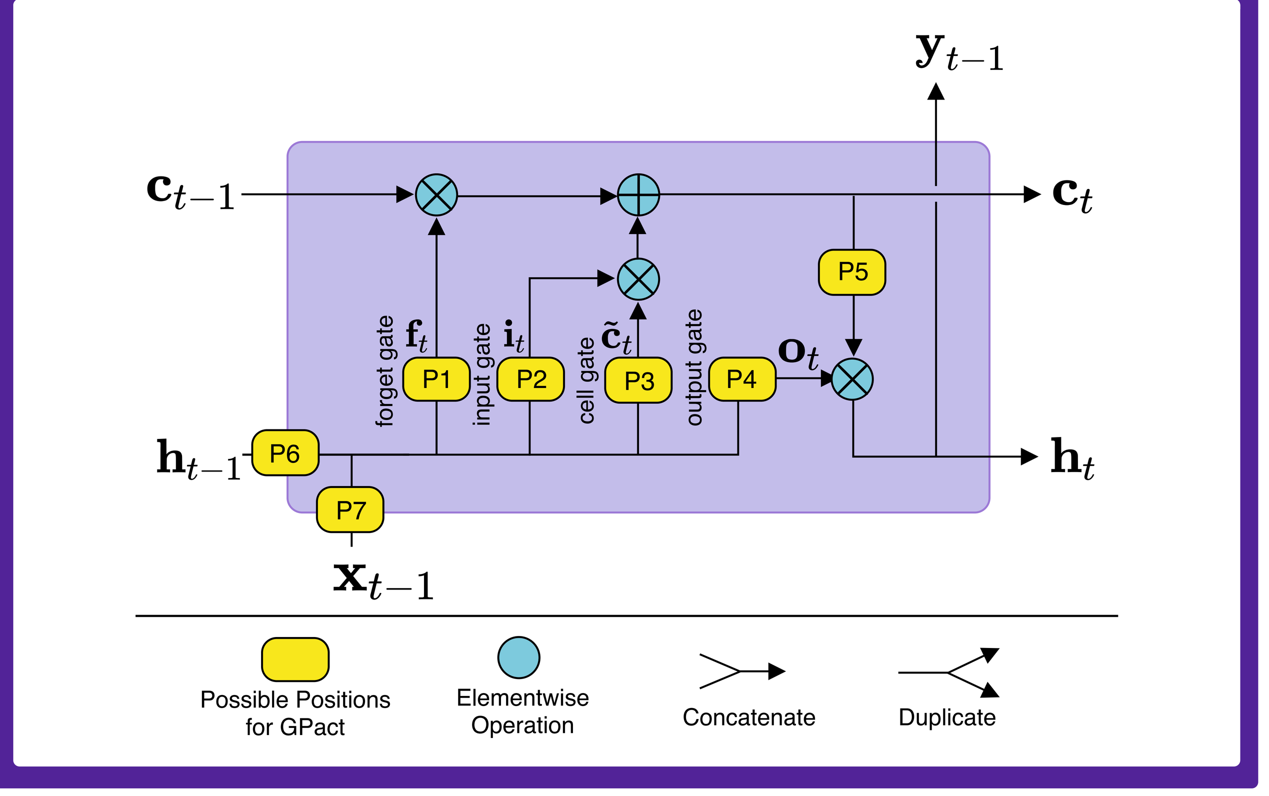
RNNLM



LSTM CELL



GP-LSTM CELL



GAUSSIAN PROCESS ACTIVATION FUNCTION

- **Standard gate:** At the d-th node, any gate can be expressed as $g_d(\mathbf{z}) = \phi(\boldsymbol{\theta}_d \bullet \mathbf{z})$, given a fixed activation function $\phi(\cdot)$ and the d-th node's weight vector $\boldsymbol{\theta}_d$.
- **Proposed gate:** Gaussian process activation function (GPact) at the d-th node is defined as

$$g_d(\mathbf{z}) = \int \sum_{k=1}^{K} \lambda_{kd}\phi_k(\boldsymbol{\theta}_d \bullet \mathbf{z})\, p(\boldsymbol{\theta}_d|\mathcal{W})\, d\boldsymbol{\theta}_d, \tag{1}$$

where $\{\lambda_{kd}\}_{k=1}^{K}$ are the coefficients for a linear combination of K *basis activation functions* $\{\phi_k(\cdot)\}_{k=1}^{K}$ and $p(\boldsymbol{\theta}_d|\mathcal{W})$ is the posterior given the observed word sequence $\mathcal{W} = < \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_T >$.

- **Variational Inference (VI):** In Bayesian inference $p(\boldsymbol{\theta}_d|\mathcal{W})$ is intractable, thus it is common to employ VI – using a learnable distribution $q_*(\boldsymbol{\theta}_d)$ to approximate $p(\boldsymbol{\theta}_d|\mathcal{W})$ with a minimal KL divergence:

$$q_*(\boldsymbol{\theta}_d) = \underset{q(\boldsymbol{\theta}_d)}{\arg\min}\, \mathrm{KL}\left\{q(\boldsymbol{\theta}_d)||p(\boldsymbol{\theta}_d|\mathcal{W})\right\} \approx \underset{\boldsymbol{\mu}_d,\boldsymbol{\Gamma}_d}{\arg\min}\, \mathrm{KL}\left\{\mathcal{N}\left(\boldsymbol{\mu}_d, \boldsymbol{\Gamma}_d^2\right)||p(\boldsymbol{\theta}_d|\mathcal{W})\right\}. \tag{2}$$

- **Upper Bound and Sampling:** The KL term in (2) is not differentiable w.r.t. $\boldsymbol{\mu}_d, \boldsymbol{\Gamma}_d$. To leverage back-propagation (BP), *KL upper bounding* and *Monte Carlo sampling* are necessary and commonly used methods to allow gradients w.r.t. $\boldsymbol{\mu}_d, \boldsymbol{\Gamma}_d$ to be calculated in a tractable way for the BP updates:

$$\mathcal{L} = -\frac{1}{S}\sum_{s=1}^{S}\sum_{t=1}^{T-1} \log P\left(\mathbf{w}_{t+1}|\mathbf{w}_t, \mathbf{h}_t; \boldsymbol{\theta}_1^{(s)}, ..., \boldsymbol{\theta}_D^{(s)}\right) + \sum_{d=1}^{D} \mathrm{KL}\{\mathcal{N}\left(\boldsymbol{\mu}_d, \boldsymbol{\Gamma}_d^2\right)||p(\boldsymbol{\theta}_d)\}, \tag{3}$$

where $\boldsymbol{\theta}_d^{(s)}$ denotes s-th sample drawn from $q(\boldsymbol{\theta}_d)$, and $p(\boldsymbol{\theta}_d) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the Gaussian prior we set.

EXPERIMENTAL SETUP

- **Tasks:** Penn Treebank (PTB) corpus, Switchboard (SWBD) and AMI meeting speech data
- **Measures:** Perplexity (PPL) for language modeling and word error rate (WER) for ASR

RESULTS

1) PPL on PTB:

Language Model	PPL	PPL(+4G)
4-gram	141.7	-
Standard LSTM	114.4	99.7
(P1) GPact as the forget gate	115.2	92.4
(P2) GPact as the input gate	115.1	91.7
(P3) GPact as the cell gate	111.9	88.3
(P4) GPact as the output gate	109.4	88.3
(P5) GPact as the $\mathbf{c}_t$ gate	111.2	88.2
(P6) GPact as a new gate for $\mathbf{h}_{t-1}$	**108.2**	**88.1**
(P7) GPact as a new gate for $\mathbf{x}_t$	112.0	90.0

2) PPL and WER on SWBD:

Language Model	PPL	WER (%)	
		swbd	callhm
4-gram	80.6	12.1	23.9
LSTM	89.3	11.4	23.9
GP-LSTM	**87.2**	**11.3**	23.9
4-gram + LSTM	71.7	11.3	23.2
4-gram + GP-LSTM	**70.1**	**11.0**	**23.1**
4-gram + LSTM + GP-LSTM	**67.2**	**10.8**	**23.0**

3) PPL and WER on AMI:

Language Model	PPL	WER (%)	
		dev	eval
4-gram	111.1	30.4	31.0
LSTM	83.4	29.4	30.0
GP-LSTM	**81.2**	**29.3**	**29.8**
4-gram + LSTM	76.8	29.3	29.8
4-gram + GP-LSTM	**74.2**	**29.0**	**29.4**
4-gram + LSTM + GP-LSTM	**71.2**	**28.7**	**29.3**

CONCLUSIONS

- GP-LSTM RNNLMs consistently showed superior results over LSTM RNNLMs in terms of both perplexity and word error rate.
- GP-LSTM RNNLMs outperformed LSTM RNNLMs in enhancing N-gram LMs.

This presentation is supported by IEEE Signal Processing Society Travel Grant for ICASSP 2019.

This research was partially funded by Research Grants Council of Hong Kong General Research grant Fund No. 14200218, and the Chinese University of Hong Kong (CUHK) grant No. 4055065.