BAYESIAN TRANSFORMER LANGUAGE MODELS FOR SPEECH RECOGNITION

Boyang Xue, Jianwei Yu, Junhao Xu, Shansong Liu, Shoukang Hu, Zi Ye, Mengzhe Geng. Xunying Liu, Helen Meng

{byxue,jwyu,jhxu,ssliu,skhu,zye,mzgeng,xyliu,hmmeng}@se.cuhk.edu.hk

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

Paper Number: 1539

1. Introduction

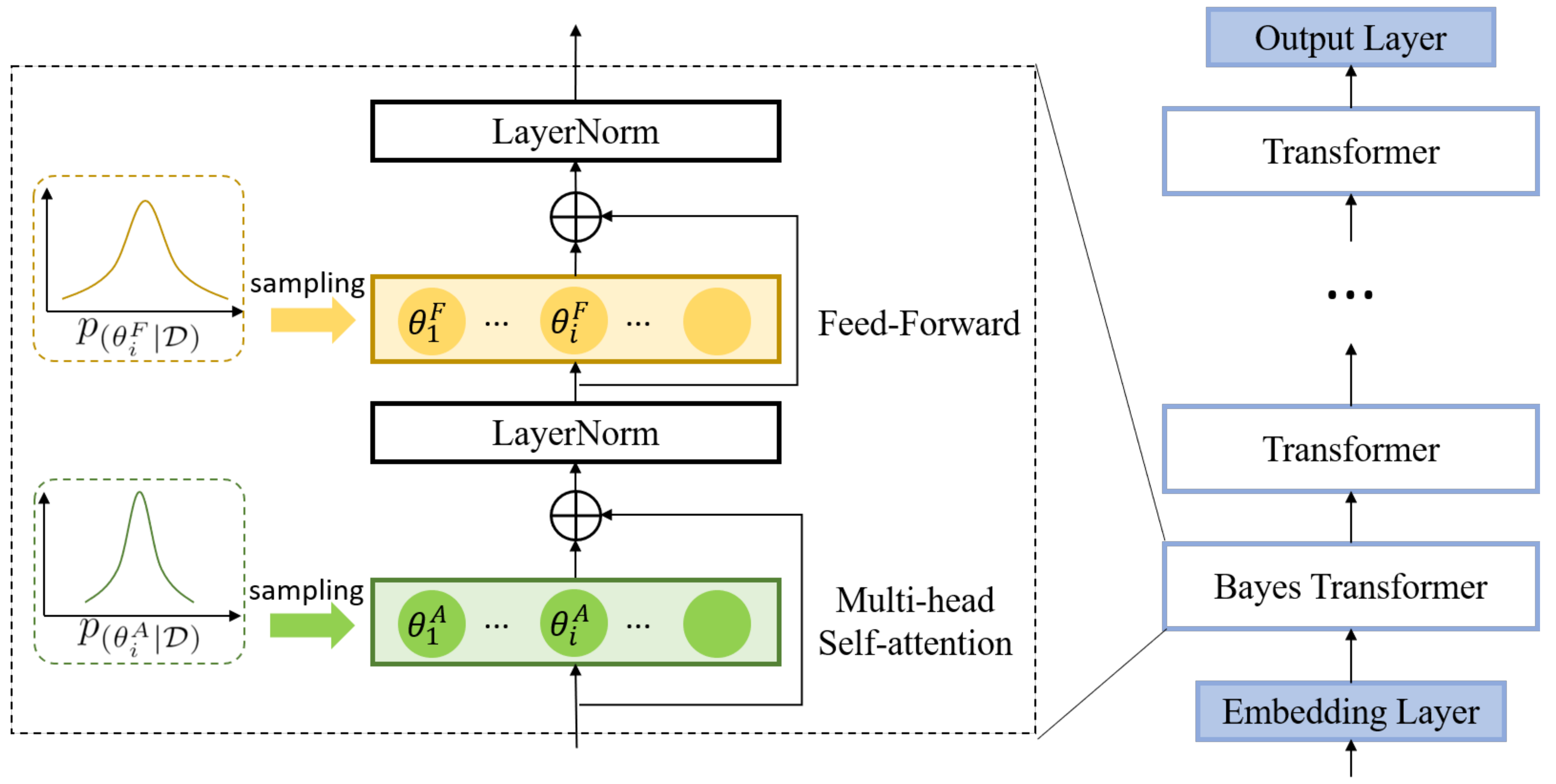
Motivation

- State-of-the-art neural language models (LMs) represented by Transformers are highly complex
- Fixed parameter estimates fail to account for model uncertainty
- Prone to over-fitting when given limited training data

Our work:

- Propose a full Bayesian learning framework to account for model uncertainty in Transformer LM estimation
- Adopt efficient variational inference based approach to estimate the latent parameter posterior distribution
- Detailed analysis on the effect of applying Bayesian estimation on different parts of Transformer LM

2. Transformer LMs



- Decoder component of Transformer architecture was adopted for LM
- Stacking of multi-head self-attention modules:

$$\boldsymbol{q}_t^i, \boldsymbol{k}_t^i, \boldsymbol{v}_t^i = \boldsymbol{Q}\boldsymbol{x}_t^{l-1}, \boldsymbol{K}\boldsymbol{x}_t^{l-1}, \boldsymbol{V}\boldsymbol{x}_t^{l-1}$$
$$\boldsymbol{h}_t^l = \left(\boldsymbol{h}_{t-1}^l, (\boldsymbol{k}_t^l, \boldsymbol{v}_t^l)\right)$$
$$\boldsymbol{y}_t^l = \boldsymbol{W}_h^l \mathrm{SelfAttention}\left(\boldsymbol{h}_t^l, \boldsymbol{q}_t^l\right) + \boldsymbol{x}_t^{l-1}$$
$$\boldsymbol{z}_t^l = \mathrm{LayerNorm}(\boldsymbol{y}_t^l)$$

- x_t^l denotes the input of the l-th Transformer block
- h_t^l stores cached key-value pairs up to word position t, enforcing left to right attention modelling over history contexts only
- Feed forward blocks following each self-attention module:

$$\boldsymbol{s}_t^l = \boldsymbol{W}_2^l GELU\left(\boldsymbol{W}_1^l z_t^l\right) + \boldsymbol{z}_t^l$$
$$\boldsymbol{x}_t^l = \mathrm{LayerNorm}(\boldsymbol{s}_t^l)$$

- For simplicity, the bias vectors are omitted in the above equations

3. Bayesian Transformer LM

- **Variational learning for Bayesian Transformer LMs:**
  - Lower bound is approximation of marginal likelihood:

$$\log P(\mathcal{D}) = \log \int P(\mathcal{D}|\boldsymbol{\Theta}) p_r(\boldsymbol{\Theta}) d\boldsymbol{\Theta}$$
$$\geq \underbrace{\sum_{n=1}^{N} \log \int P(W^n|\boldsymbol{\Theta}) q(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}_{\mathcal{L}_1} - \underbrace{KL(q(\boldsymbol{\Theta})||p_r(\boldsymbol{\Theta}))}_{\mathcal{L}_2} = \mathcal{L}$$

  - $\mathcal{D}$ represents the whole training set for model development
  - $q(\boldsymbol{\Theta})$ denotes the variational approximation of parameter posterior distribution $p(\boldsymbol{\Theta}|\mathcal{D})$
  - $p_r(\boldsymbol{\Theta})$ denotes the prior distribution of $\boldsymbol{\Theta}$
  - $q(\boldsymbol{\Theta})$ and $p_r(\boldsymbol{\Theta})$ assumed to be **diagonal Gaussian**

$$q(\boldsymbol{\Theta}) \sim N(\boldsymbol{\Theta}; \boldsymbol{\mu}, \boldsymbol{\sigma}), \qquad p_r(\boldsymbol{\Theta}; \boldsymbol{\mu}_r, \boldsymbol{\sigma}_r)$$

  - Allowing KL term to be in a differentiable close form
  - Monte Carlo sampling used to approximate the marginal likelihood $\mathcal{L}_1$:

$$\mathcal{L} \approx -KL(q(\boldsymbol{\Theta})||p_r(\boldsymbol{\Theta})) + \frac{1}{K}\sum_{k=1}^{K} \log P(\boldsymbol{W}|\boldsymbol{\Theta}_k)$$

  - With re-parameterization used when sampling $\boldsymbol{\Theta}_k$

$$\boldsymbol{\Theta}_k = \boldsymbol{\mu} + \boldsymbol{\epsilon_k} \odot \boldsymbol{\sigma}, \qquad \boldsymbol{\epsilon_k} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

  - Estimation of variational distribution parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$ integrated with SGD based back propagation

$$\frac{\partial \mathcal{L}}{\partial \mu_i} = \frac{1}{K}\sum_{k=1}^{K} \frac{\partial \mathcal{L}_1}{\partial \mu_i} - \frac{\mu_i - u_{r,i}}{\sigma_i^2}$$
$$\frac{\partial \mathcal{L}}{\partial \sigma_i} = \frac{1}{K}\sum_{k=1}^{K} \frac{\partial \mathcal{L}_1}{\partial \sigma_i} - \frac{\sigma_i^2 - \sigma_{r,i}^2}{\sigma_i^2}$$

- **Implementation details**
  - Applying Bayesian estimation on part of the model parameters
  - Parameters obtained from standard Transformer LM is used as the prior's mean $\boldsymbol{\mu}_r$, prior's variance is set to be 1
  - Only use the mean of the Bayesian parameters in evaluation

$$P(\boldsymbol{w_t}|\boldsymbol{w_1}, ..\boldsymbol{w_{t-1}}) = \int P(\boldsymbol{w_t}|\boldsymbol{w_1}, ..\boldsymbol{w_{t-1}}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}|\mathcal{D}) d\boldsymbol{\Theta}$$
$$\approx P(\boldsymbol{w_t}|\boldsymbol{w_1}, ..\boldsymbol{w_{t-1}}, \boldsymbol{\Theta}_{mean})$$

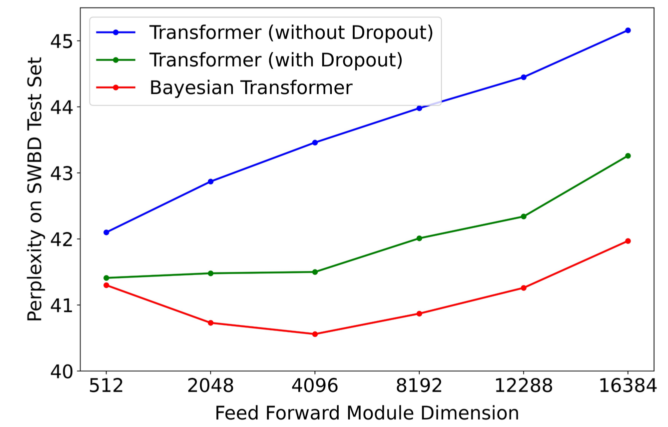
4. Experiments & Results

Experiments on Conversational Telephone Speech

- Datasets: 300 hour Switchboard for acoustic modelling; 34M words of Switchboard+Fisher transcriptions for language modelling; 30k vocabulary lexicon.
- Acoustic model: TDNN-F based hybrid model featuring speech perturbation, i-Vector, LHUC speaker adaptation and (LF-MMI) sequence training

ID	LM	Bayesian		PPL	eval2000		rt02			rt03	
		Block	Position	(swbd)	swbd	callhm	swbd1	swbd2	swbd3	fsh	swbd
1	4gram	Not Applied		-	9.7	18.0	11.5	15.3	20.0	12.6	19.5
2	Transformer(+4g)	Not Applied		41.50	7.9	15.7	9.5	12.8	17.4	10.4	17.3
3	Bayes Transformer(+4g)	-	EMB	41.01	7.7	15.6	9.5	12.6	17.1†	10.2	17.1†
4		1	MHA	40.95	7.7	15.5	9.5	12.5†	17.1†	10.2	17.1†
5		1	FF	40.65	7.7	15.4†	9.4	12.6†	17.0†	10.2†	17.0†
6		1-2	FF	41.11	7.7	15.6	9.5	12.6	17.2	10.3	17.1
7		1-3	FF	42.45	7.8	15.8	9.5	12.7	17.2	10.3	17.2
8		1-4	FF	47.54	8.0	16.0	9.9	13.0	17.6	10.7	17.5
9		1-5	FF	54.19	8.3	16.2	10.2	13.5	18.0	11.1	18.0
10		1-6	FF	74.50	8.9	17.3	10.8	14.3	18.7	12.0	18.8
11	+Transformer(+4g)	-	EMB	40.03	7.7	15.5	9.4	12.6†	17.1†	**10.1**†	17.0†
12		1	MHA	39.70	**7.6**†	15.4†	9.3	**12.5**†	**17.0**†	**10.1**†	**16.9**†
13		1	FF	**39.42**	**7.6**†	**15.2**†	9.3	**12.5**†	**17.0**†	**10.1**†	**16.9**†

- Proposed Bayesian Transformer LMs (line 11-13) outperform the baseline Transformer LM(line 2) in terms of both PPL and WER by statistically significant margin from 0.3% to 0.5% absolutely
- Applying Bayesian estimation on the feed forward (FF) module outperforms using Bayesian estimation on multi-head self-attention (MHA) or embedding (EMB) layer
- Compared with applying Bayesian estimation to multiple Transformer blocks (line 6-10), adopting Bayesian estimation on the lowest Transformer block (line 5) produced the best PPL and WER



ID	LMs	Adapt	PPL	WER(%)
1	4gram	✗	17.07	30.67
2	Transformer(+4g)	✗	21.83	30.65
3		fine-tuning	14.56	30.25
4	Bayes Transformer(+4g)	✗	19.88	30.49
5		bayes-adapt	13.99	**29.88**†

- Performance improvements consistently obtained on a cross domain LM adaptation task requiring porting a Transformer LM trained on the Switchboard and Fisher data to a low-resource DementiaBank elderly speech corpus.

5. Conclusions

- The proposed Bayesian learning framework can improve the performance and robustness of Transformer LMs in both model training and adaptation.
- The parameters associated with the higher Transformer blocks are expected to be more deterministic than those experienced in the lower

- This research is supported by Hong Kong RGC GRF grant No. 14200218, 14200220, Theme-based Research Scheme T45-407/19N, Innovation & Technology Fund grant No. ITS/254/19, and Shun Hing Institute of Advanced Engineering grant No. MMT-p1-19.