Interview Presentation

XUE Boyang，薛博阳
The Chinese University of Hong Kong

Sep 21th, 2022

Presentation Outline

- Profile - Personal Information.
- Projects - Three Research Experiences.
- Prospects - Viewing NLP Research.

Profile

Name: XUE, Boyang.

Research Interest: Language Modeling, Speech Recognition, Machine Learning.

Education:

Bachelor Degree in Huazhong University of Science and Technology (HUST). Sep. 2016 - Jun. 2020

Faculty: Automation, Excellent Class. (GPA 3.54/4.00, Ranking: 7/30)

Research Assistant in The Chinese University of Hong Kong (CUHK). Sep. 2020 - Jul. 2021

Human-Computer Communications Lab, System Engineering Department.

Second-year PhD Candidate in The Chinese University of Hong Kong (CUHK). Aug. 2021 - Present

Human-Computer Communications Lab, System Engineering Department. (GPA 3.79/4.00)

Profile

Publications (First Author):

Bayesian Neural Network Language Modeling for Speech Recognition. in *IEEE/ACM TASLP, 2022.*

Bayesian Transformer Language Models for Speech Recognition. in *IEEE ICASSP, 2021.*

Deep Learning based Patient-Specific Fetal Heart Rate Detection System on FECG. *Chinese Patent, 2020.*

Awards & Honors:

National Grand Prize in the 14th NXP Cup National University Intelligent Car Race (Top 3/468).

Excellent Graduate (Top 15%), University Scholarship (2018, Top 10%).

English: IELTS – L: 6.5, R: 7.5, W: 7.0, S: 5.5, Overall: 6.5.

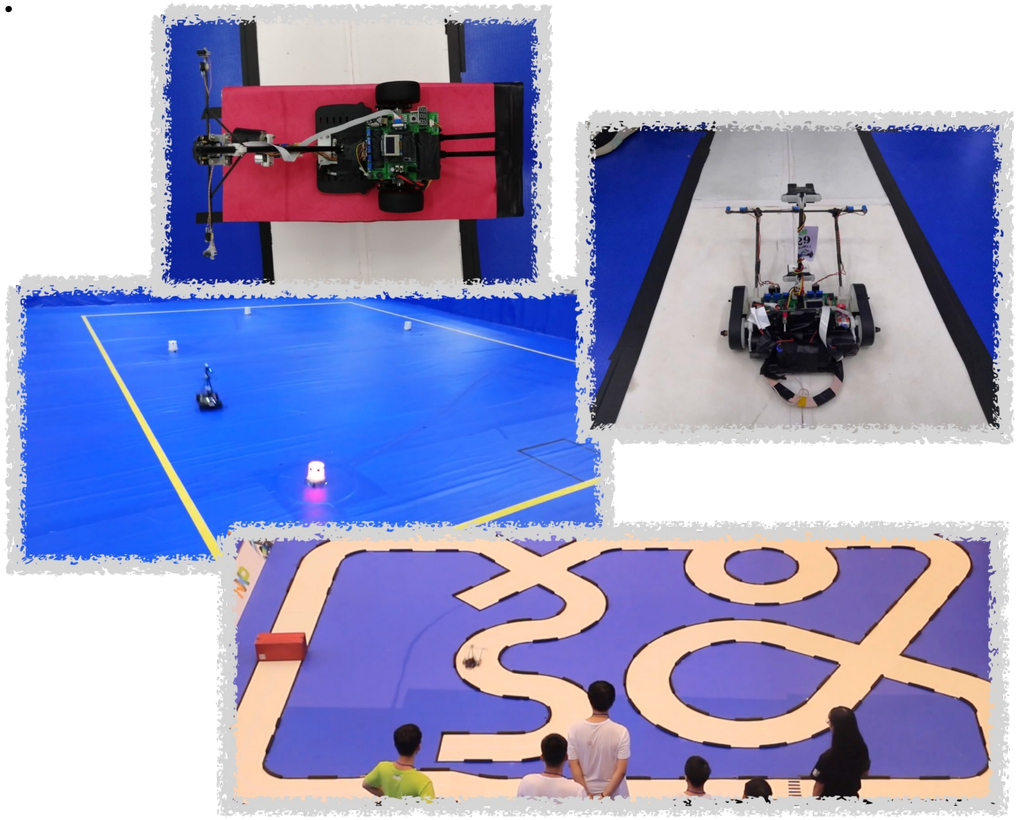
Programming and Development: C, Python, MATLAB, Linux Shell, Latex, PyTorch, Kaldi, et al.

Others: Served as a part-time author in PaperWeekly and Zhihu; Responsible for a PRML project in Datawhale.

Projects - 1

NXP Cup National University Intelligent Car Race

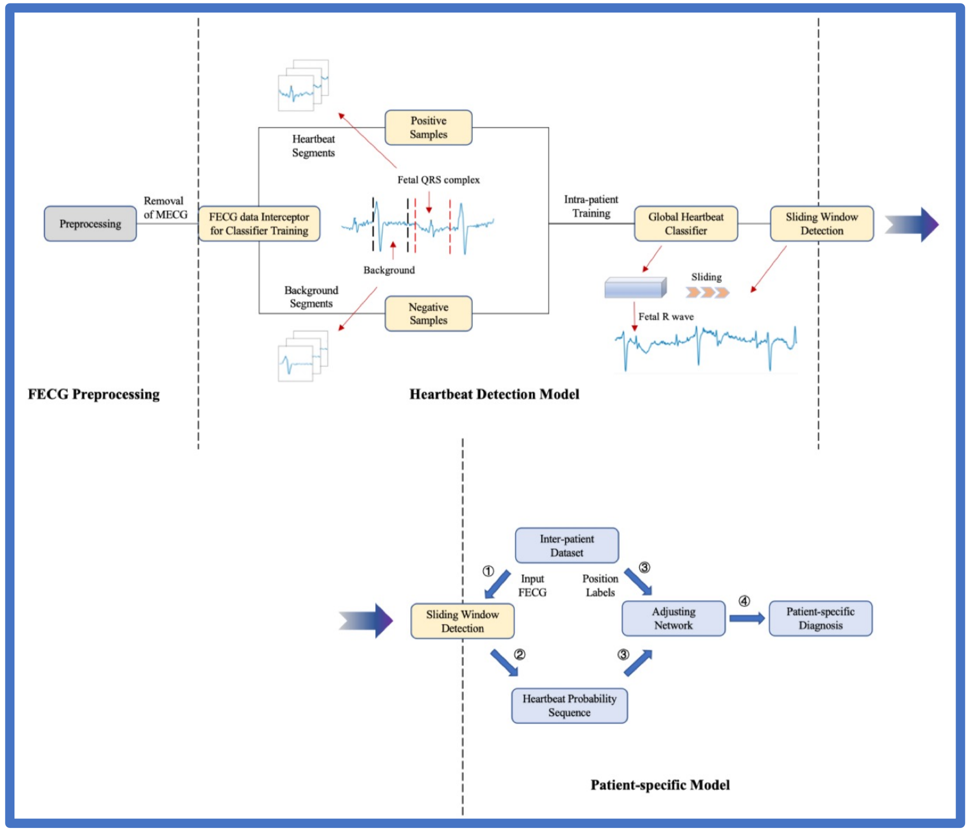
- **Dec.2017 - Aug.2019**, Team Leader in Robotic Team, HUST.
- **Group1**: Kalman filter based self-balanced car.
- **Group2**: Tracking car for beacon destination.
- **Group3**: Wireless charge based energy-efficient car.
- **Responsible** for programming on Cortex-M chips.
  - Electromagnetic signal and image processing.
  - Motion control using PID algorithm.
- **Auxiliary** for circuits design and mechanics.



- **Grand Prize (Top 3/468)** in 14th NXP Cup National University Intelligent Car Race.

Deep Learning based Patient-Specific FHR Detection System

- **Sep.2019 - Jun.2020**, Research Intern in Intelligent Manufacturing and Data Science Lab, HUST.
- **Co-operated** with Tongji Hospital, HUST.
- **Main Contributions:**
  - Improvements of non-invasive fetal heart rate detection on FECG.
  - CNN-LSTM based model to detect fetal QRS wave on FECG.
  - Patient-specific detection method to alleviate intra-differences.
- **Dataset:** PhysioNet FECG and Tongji FECG dataset.
- **Clinical Value:** Prenatal diagnosis to reduce fetal mortality.
- **Chinese Patent** Granted in 2020.



Projects - 3

Bayesian Learning based Neural Network Language Models

- **Sep.2020 - May.2022**, Research Assistant and PhD Student, CUHK.
- **Language Model Application Examples (Key component of a range of Natural Language Processing tasks):**
  - Speech Recognition: $P\,(\text{read a book}) > P\,(\text{read a boot})$; Machine Translation: $P\,(\text{what is this}) > P\,(\text{this is what})$.
- **Developments of Statistical Language Models:**
  - N-gram LMs → Neural Network LMs (FNN, RNN, LSTM, Transformer) → Pre-trained LMs (BERT, GPTs)
- **Motivation:** The use of **point estimated parameters** of **highly complex** neural LMs fails to account for model uncertainty and is prone to overfitting and poor generalization on limited training data.
- **Regularization Methods:** L1 & L2 penalty and MAP estimation; Dropout method.
- **Address:** A Bayesian estimated neural network LMs for uncertainty modeling.

Bayesian Learning Framework and Variational Inference

- **Bayesian Learning Framework**
  - Bayesian Neural Network (BNN) LMs: uncertainty modeling on parameters. $\log P(W|\mathcal{D}) = \log \int P(W|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\mathcal{D})\mathrm{d}\boldsymbol{\Theta}$
  - Gaussian Process (GP) based NNLMs: uncertainty modeling on both parameters and structures.
    - Space viewed GP: $f(\boldsymbol{x}) = \boldsymbol{\lambda}^{\mathrm{T}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{j=1}^{K} \lambda^j \phi^j(\boldsymbol{x})$ $\quad \log P(W|\mathcal{D}) = \log \iint P(W|\boldsymbol{\Theta}, \boldsymbol{\lambda})p(\boldsymbol{\Theta}|\mathcal{D})p(\boldsymbol{\lambda}|\mathcal{D})\mathrm{d}\boldsymbol{\Theta}\mathrm{d}\boldsymbol{\lambda}$
  - Variational Neural Network (VNN) LMs: uncertainty modeling on hidden representations.

$$\log P(W|\mathcal{D}) = \log \prod_{t=1}^{n} P(w_t|w_0, \dots, w_{t-1}, \mathcal{D}) \approx \log \prod_{t=1}^{n} \int P(w_t|w_0, \dots, w_{t-1}, z_t)p(z_t|h_t, \mathcal{D})\mathrm{d}z_t$$

- **Variational Inference for Bayesian NNLMs:**
  - Variational lower bound maximization; Approximate $p(\boldsymbol{\Theta}|\mathcal{D})$ with $q(\boldsymbol{\Theta})$ and $p_{\mathrm{r}}(\boldsymbol{\Theta})$.

$$\log P(\mathcal{D}) = \log \int P(\mathcal{D}|\boldsymbol{\Theta})p_{\mathrm{r}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \geq \int \log P(\mathcal{D}|\boldsymbol{\Theta})q(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} - \mathrm{KL}[q(\boldsymbol{\Theta})||p_{\mathrm{r}}(\boldsymbol{\Theta})]$$

  - Three Tricks: 1) Gaussian Assumption; $q(\boldsymbol{\Theta})\sim\mathcal{N}(\boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, $\quad p_{\mathrm{r}}(\boldsymbol{\Theta})\sim\mathcal{N}(\boldsymbol{\Theta}|\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r^2)$

    2) Mento Carlo Sampling; 3) Re-parameterization Trick for Sampling. $\boldsymbol{\Theta} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}$, $\quad \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0}, \mathbf{1})$

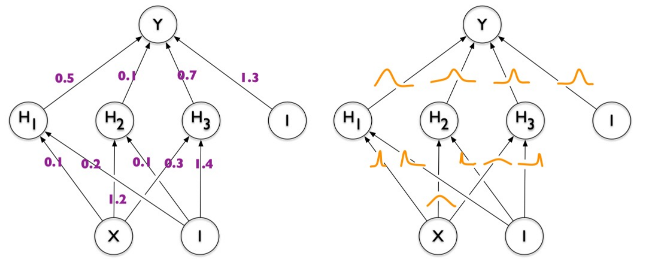


Fig. 1 Point and Bayesian estimated Neural Networks

Scalability Issue and Uncertainty Analysis of Bayesian NNLMs

- **Increased Computational Cost :**
  - linearly in terms of Monte Carlo samples - A minimal number for balance between convergence speed and performance.
  - exponentially with respect to number of positions to be Bayesian estimated - NAS to automatically select.

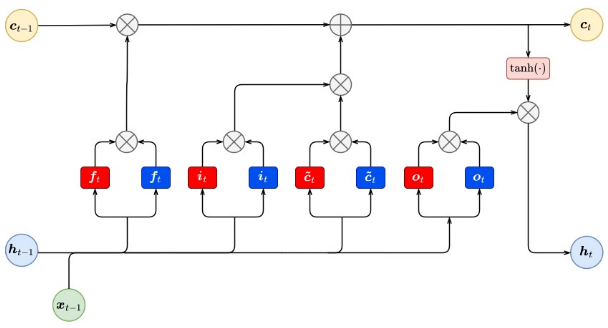


Fig. 1

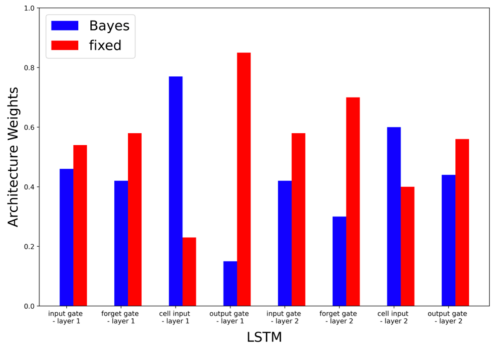


Fig. 2

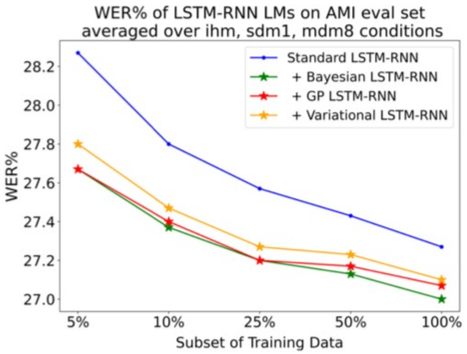


Fig. 3

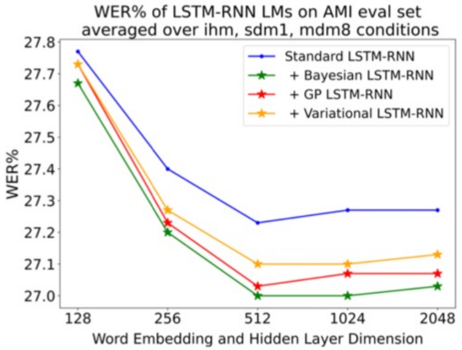


Fig. 4

- **Uncertainty Analysis:**
  - Fixed model size varying from different sizes of data.
  - Changing model complexity versus constant data quantity.
  - Parameter uncertainty: Signal-to-Noise Ratio (SNR). $\mathbf{SNR}_{\Theta} = \frac{|\mu|}{\sigma}$

ID	LM	Subset of Training Data / SNR				
1	B-LSTM	5%/**0.3**	10%/**0.4**	25%/**0.5**	50%/**0.7**	100%/**1.1**
2	B-Transformer	5%/**0.6**	10%/**0.8**	25%/**1.1**	50%/**2.0**	100%/**3.3**
	LM	**Hidden (FFN) Layer Dimensionality / SNR**				
3	B-LSTM	128/**5.1**	256/**2.6**	512/**1.4**	1024/**1.1**	2048/**2.0**
4	B-Transformer	512/**8.2**	1024/**4.3**	2048/**3.0**	4096/**3.3**	8192/**3.7**

Experimental Results and Developments of Bayesian NNLMs

- **Bayesian Transformer Language Models for Speech Recognition. in *IEEE ICASSP, 2021.***
  - **Dataset:** 34M Switchboard Telephone corpus; 2.4M DementiaBank Pitt corpus.

ID	LM	PPL (swbd)	eval2000 swbd	eval2000 callhm	rt02 swbd1	rt02 swbd2	rt02 swbd3	rt03 fsh	rt03 swbd
1	4gram	-	9.7	18.0	11.5	15.3	20.0	12.6	19.5
2	Standard Transformer	41.50	7.9	15.7	9.5	12.8	17.4	10.4	17.3
3	Bayesian Transformer	**39.42**	**7.6**	**15.2**	**9.3**	**12.5**	**17.0**	**10.1**	**16.9**

ID	LMs	Adapt	PPL	WER(%)
1	Standard Transformer	fine-tuning	14.56	30.25
2	Bayesian Transformer	bayes-adapt	13.99	**29.88**

  - **Main Contributions:** 1) First Attempt of variational BNN based Transformer LM; 2) Domain adaptation.

- **Bayesian Neural Network Language Modeling for Speech Recognition. in *IEEE/ACM TASLP, 2022.***
  - **Dataset:** 15M AMI Meeting Room data; 2.5M LRS2 Overlapped Speech corpus.

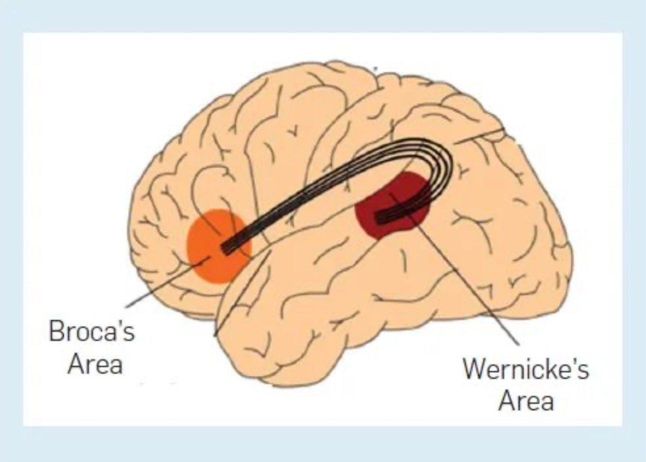
ID	LMs	PPL (eval)	WER(%) of eval ihm	sdm1	mdm8	Avg.
1	Transformer + LSTM + 3g	55.1	15.6	34.3	30.4	26.8
2	Transformer(L2) + LSTM(L2) + Transformer + LSTM + 3g	51.4	15.5	34.2	30.2	26.6
3	Bayesian Transformer + Bayesian LSTM + Transformer + LSTM + 3g	47.9	**15.3**†	**33.8**†	**29.9**†	**26.3**†

ID	LM	PPL (Test)	WER(%) clean	TF masking	Filter & Sum	MVDR	Avg.
1	Transformer + LSTM + 4g	65.8	5.3	12.2	11.3	11.6	10.1
2	Transformer(L2) + LSTM(L2) + Transformer + LSTM + 4g	64.8	5.2	12.2	11.1	11.4	10.0
3	GP Transformer + GP LSTM + Transformer + LSTM + 4g	63.7	**4.7***	**11.4***	**10.7***	**10.8***	**9.4***

  - **Main Contributions: 1)** Systematic Bayesian framework (BNN, GP, VNN) on both Transformer and LSTM-RNN LMs; **2)** Multiple LM combinations for SOTA performance; **3)** Ablation study of L1, L2 and MAP regularizations; **4)** Computational cost reduction; **5)** Uncertainty analysis in terms of model complexity and data quantity.

Some Perspectives in Language Modeling Research

- Language modeling is the key to make the machine understand humans and achieve artificial intelligence.
- Model compression and quantization - for both model size and computational cost.
- Extracting more syntactic representations combined with semantic information in LMs.
  - Brain science and cognitive science (Broca's Area and Wernicke's Area).
- Multi-modal language modeling to mimic human's behaviors.
- PhD Pursuit: broaden my horizon for more interested and original research.



Thank you !