LEARNING SPARSE STRUCTURED ENSEMBLES WITH STOCASTIC GTADIENT MCMC SAMPLING AND NETWORK PRUNING

Yichi Zhang

Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab
Department of Electronic Engineering
Tsinghua University, Beijing, China

September 19, 2018

Outline

- Motivation & Problem
- Related Work
- Our solution: Mix of multiple ingredients
  - Learning ensembles via SG-MCMC sampling
  - Cost reduction via structured model compression
  - Experimental results
- Conclusion & Future Work

Ensemble of Neural Networks

- Ensemble models are a **group of models** that work collectively to get the averaged prediction.

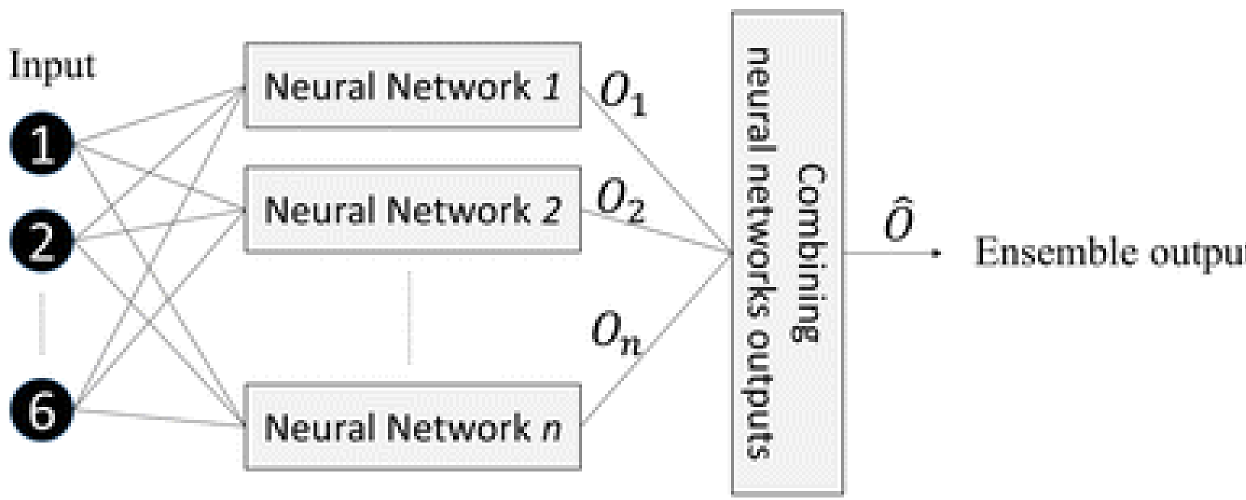


Figure from Effat Dehghanian et al. 2015

Ensemble of Neural Networks

- Ensemble gives a **great boost in accuracy** because it does not rely on a single model for prediction.

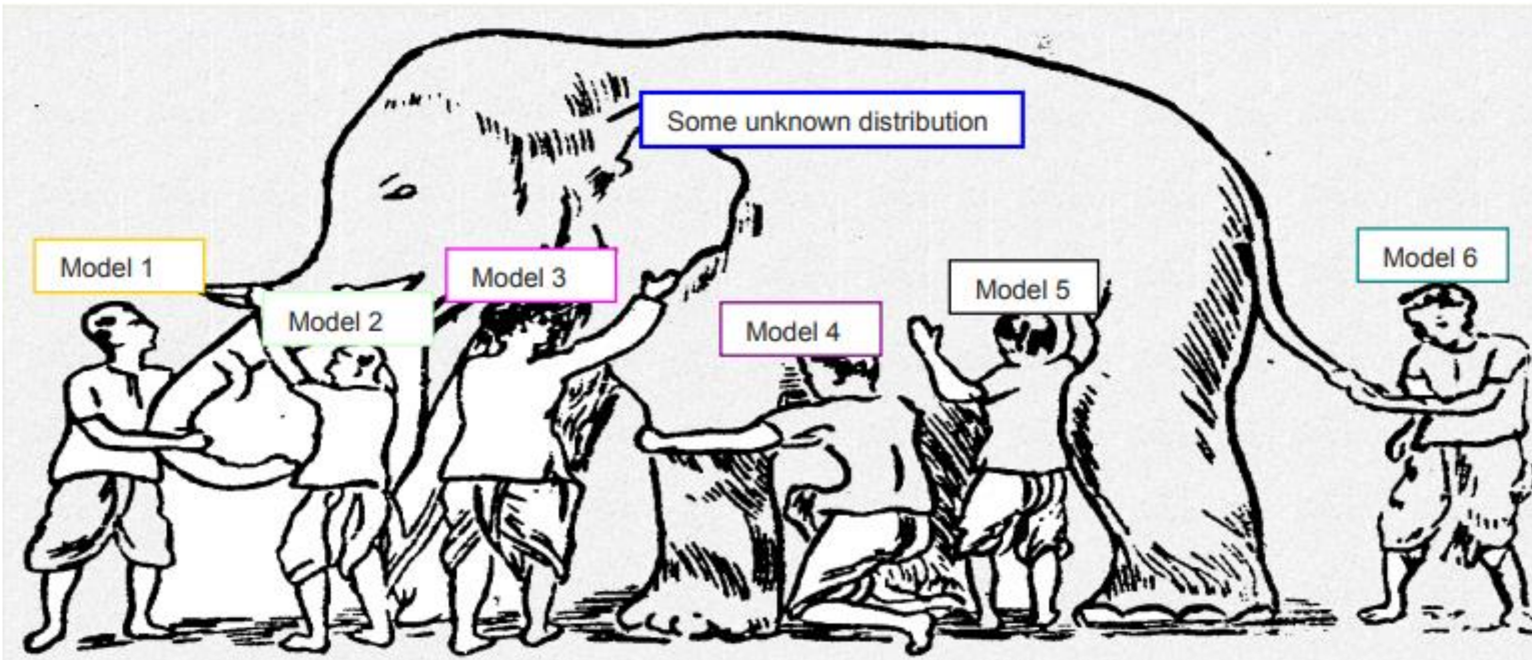


Figure from Alejandro Correa et al. 2013

Ordered by classification error

Team name	Entry description
WMW	Ensemble C [No bounding box results]
WMW	Ensemble E [No bounding box results]
WMW	Ensemble A [No bounding box results]
WMW	Ensemble D [No bounding box results]
WMW	Ensemble B [No bounding box results]
Trimps-Soushen	Result-1
Trimps-Soushen	Result-2
Trimps-Soushen	Result-3
Trimps-Soushen	Result-4
Trimps-Soushen	Result-5
NUS-Qihoo_DPNs (CLS-LOC)	[E2] CLS:: D

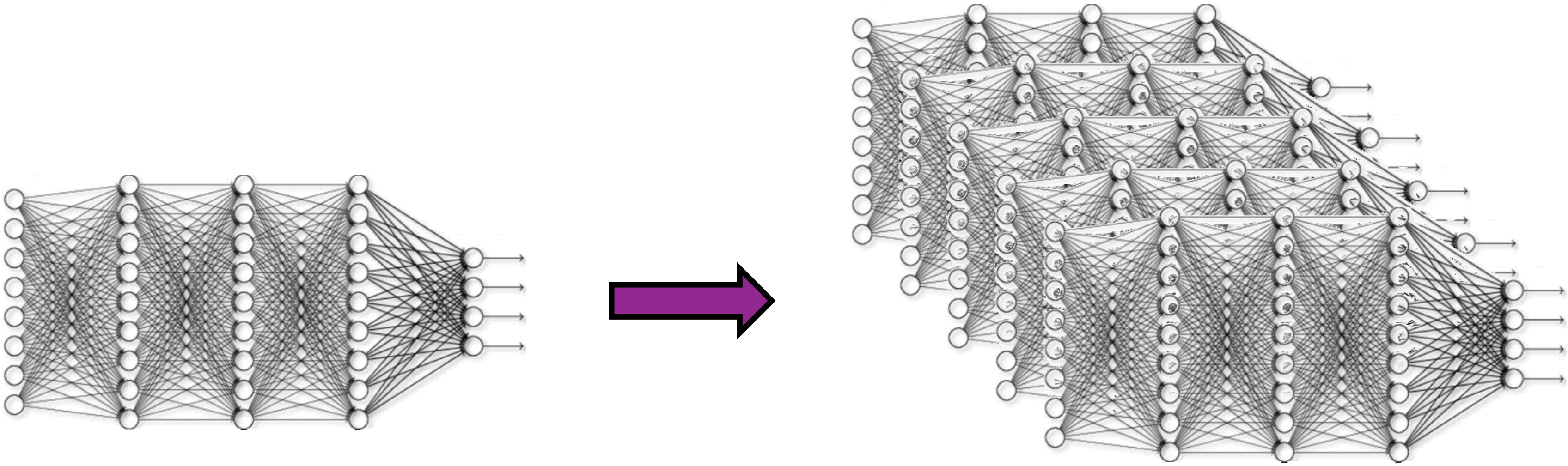
WER test-clean	Paper	Published	Notes
5.83%	Deep Speech 2: End-to-End Speech Recognition in English and Mandarin	December 2015	*Humans*
3.19%	The CAPIO 2017 Conversational Speech Recognition System	April 2018	TDNN + TDNN-LSTM + CNN-bLSTM + Dense TDNN-LSTM across two kinds of trees
[illegible]	Improved training of end-to-end [illegible]	Interspeech, Sept 2018	encoder-attention-decoder end-to-end model

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Model	EM	F1
Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
MARS (ensemble) *YUANFUDAO research NLP*	**83.982**	**89.796**
QANet (ensemble) *Google Brain & CMU*	83.877	89.737
Hybrid AoA Reader (ensemble) *Laboratory of HIT and iFLYTEK Research*	82.482	89.281
MARS (single model) *YUANFUDAO research NLP*	83.122	89.224
QANet (single) *Google Brain & CMU*	82.471	89.306
QANet (ensemble) *Google Brain & CMU*	82.744	89.045

CLASSIFIER	
large conv. net, unsup featu[illegible]	
large conv. net, unsup pretraining [no distortions]	0.60
large conv. net, unsup pretraining [elastic distortions]	0.39
large conv. net, unsup pretraining [no distortions]	0.53
large/deep conv. net, 1-20-40-60-80-100-120-120-10 [elastic distortions]	0.35
committee of 7 conv. net, 1-20-P-40-P-150-10 [elastic distortions]	0.27 +-0.02
committee of 35 conv. net, 1-20-P-40-P-150-10 [elastic distortions]	0.23

Problems



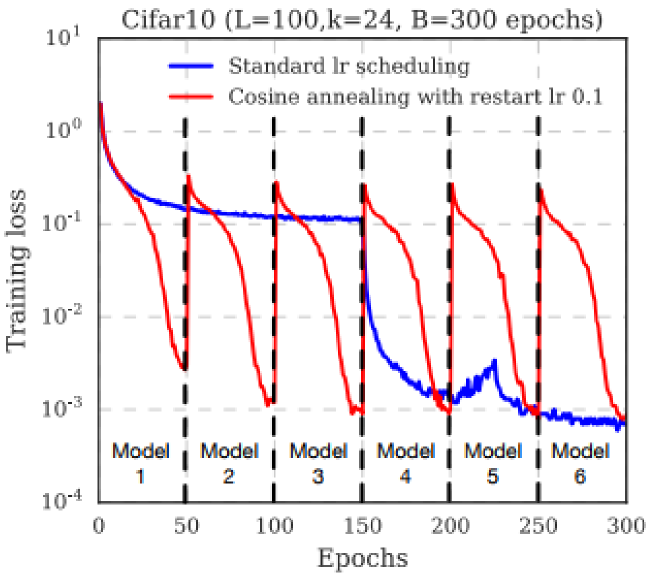
N networks

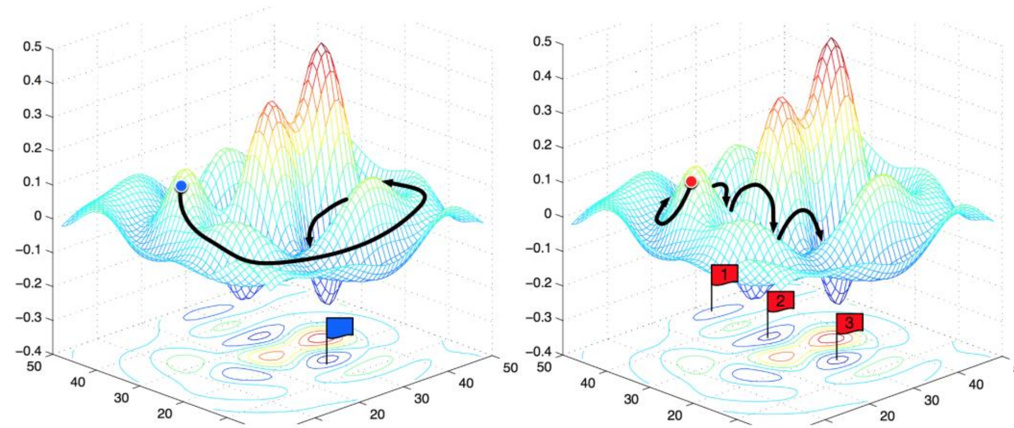
- **Training problem:** N times training time
- **Testing problem:** N times memory/testing time cost

Related Work

Snapshot ensembles: Train 1, get m for free (Gao Huang et al. 2017)

- Obtain multiple snapshot models within a single training process.
- Empirical cyclic learning rate settings.





Related Work

- The recent progress in Bayesian posterior sampling:
  Stochastic Gradient Markov Chain Monte Carlo sampling algorithms
  (Max Welling et al. 2011, Tianqi Chen et al. 2014, Zhe Gan et al. 2016)

- SG-MCMC works by adding a scaled gradient noise to Stochastic optimization method which is proved to have the following benefits :
  (i) Theoretically interpretable
  (ii) Efficient exploration of the model parameter space
  (iii) Scalable and simple

Related Work

- Testing problem: N times memory/testing time cost

Model compression via
Network pruning and retraining
(Song Han et al. 2015, 2017).

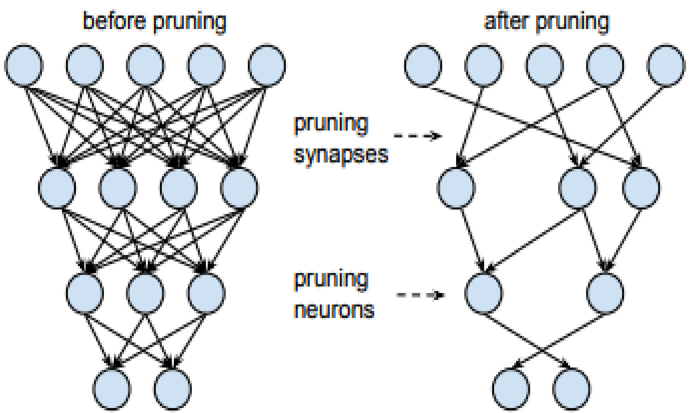


Figure from Song Han et al. 2015

Sparse structure learning via
Group Lasso penalty (Ming Yuan et al. 2006)
on deep models (Wei Wen et al. 2016, 2017).

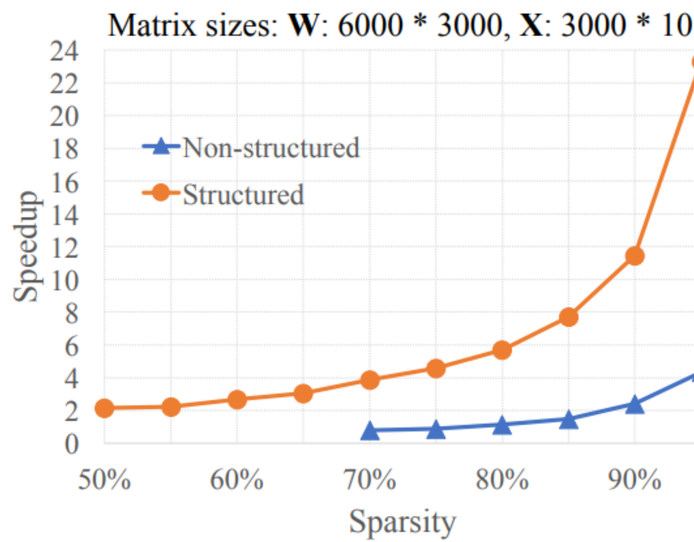


Figure from Wei Wen et al. 2017

Our Propose

SG-MCMC based Bayesian learning + **Group Sparse Prior** + **Network Pruning & Retraining**

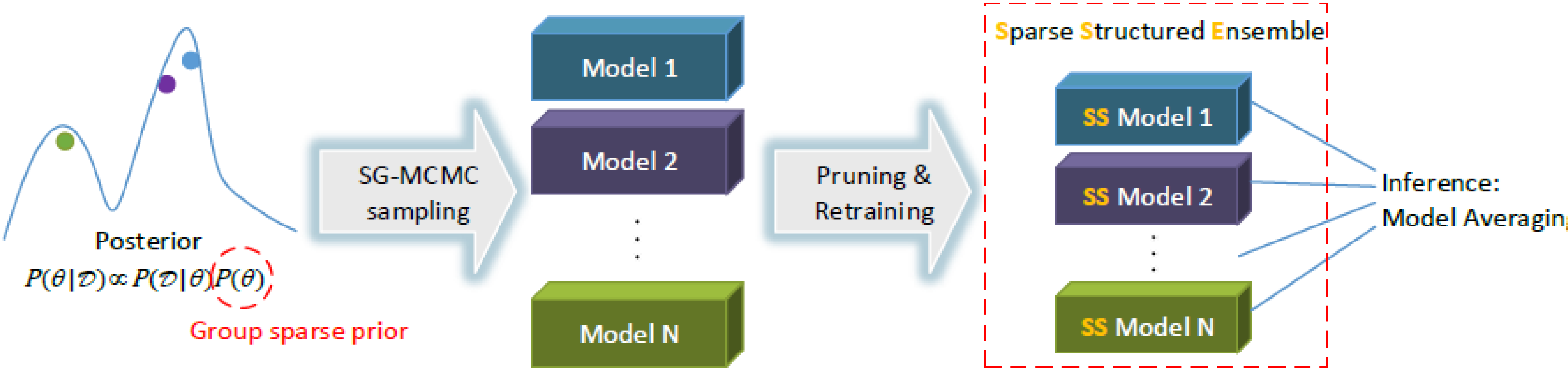


Figure 1: Overview of our two-stage method for learning SSEs.

Bayesian Neural Network Framework

- Denote θ as all the trainable parameters in a neural network.
- Given data $D = \{(x_i, y_i)\}_{i=1}^{N}$, where input $x_{\boldsymbol{i}} \in \mathbb{R}^D$ and label $y_i \in \mathcal{Y}$
- The goal of training is to evaluate the posterior distribution:

$$p(\theta|D) \propto p(\theta) \prod_{i=1}^{N} p(y_i|x_i, \theta) \qquad (1)$$

- Given a testing input $\tilde{x}$, the Bayesian predictive distribution

$$p(\tilde{y}|\tilde{x}, D) = \mathrm{E}_{p(\theta|D)}[p(\tilde{y}|\tilde{x}, \theta)] = \int_{\theta} p(\tilde{y}|\tilde{x}, \theta) p(\theta|D) d\theta \qquad (2)$$

$$p(\tilde{y}|\tilde{x}, D) \approx \frac{1}{M} \sum_{m=1}^{M} p(\tilde{y}|\tilde{x}, \theta_m) \quad ,\theta_m \sim p(\theta|D) \qquad (3)$$

can be considered as the average of NN softmax outputs.

Training: SG-MCMC Sampling

- Goal: sample $\theta \sim p(\theta|D)$, obtain $\{\theta_m\}_{m=1}^{M}$
- Method: Stochastic Gradient Markov Chain Monte Carlo (SG-MCMC)

Stochastic Gradient Descent:

$$\tilde{g}_t = \frac{N}{n}\sum_{i=1}^{n} \nabla \log p\left(y_t^{(i)} \middle| x_t^{(i)}, \theta_t\right),$$

$$\Delta\theta_t = \epsilon_t \tilde{g}_t$$

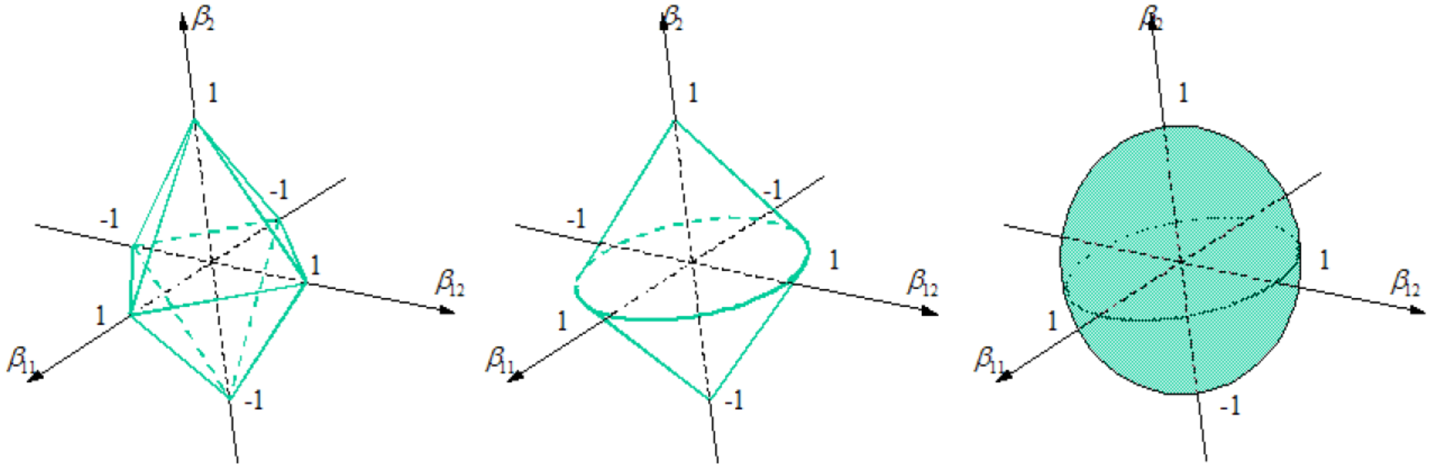
Stochastic Gradient Langevin Dynamic (Max Welling and Yee W Teh, 2011):

$$\tilde{g}_t = \nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p\left(y_t^{(i)} \middle| x_t^{(i)}, \theta_t\right),$$

$$\Delta\theta_t = \epsilon_t \tilde{g}_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, 2\epsilon_t)$$

Group Sparse Prior

L_1: *Random sparsity* | L_{21}: *Group sparsity* | L_2: *no sparsity*



$\beta_{11} = 0 \; or$
$\beta_{12} = 0 \; or$
$\beta_2 = 0$

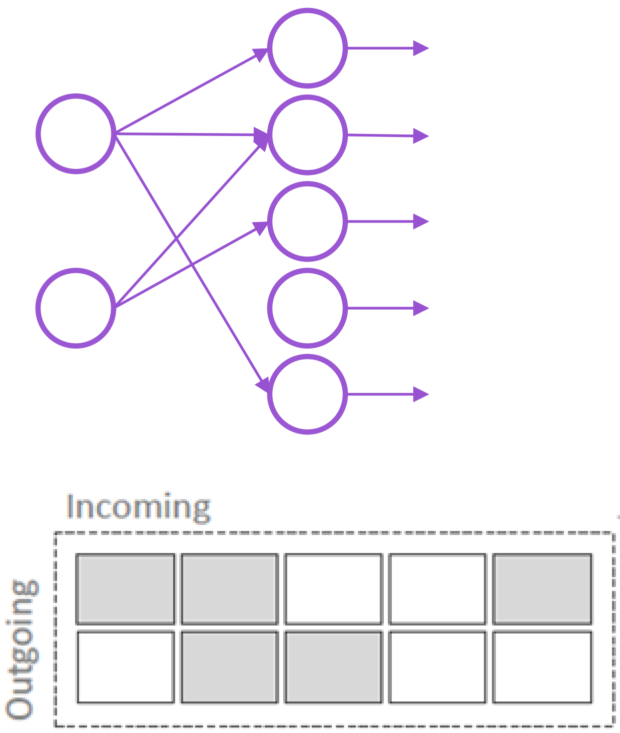
$(\beta_{11}, \beta_{12}) = 0 \; or$
$\beta_2 = 0$

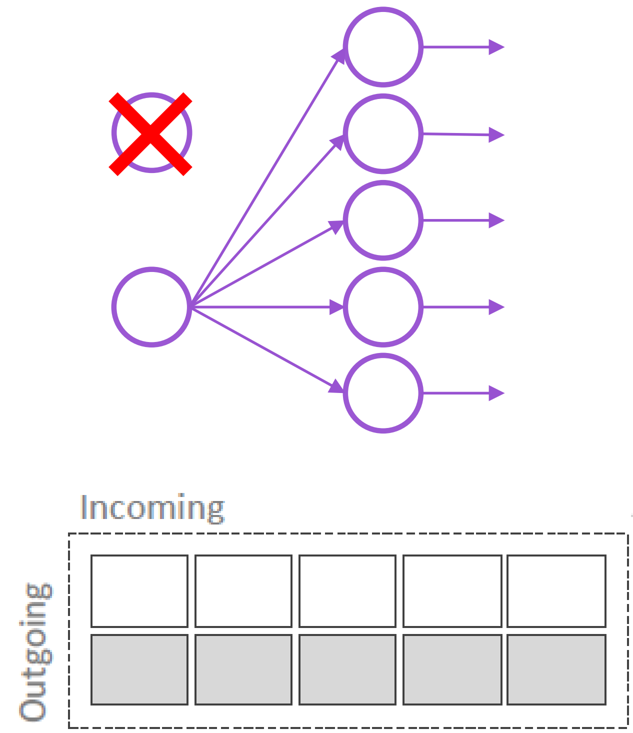
group sparse prior

Figure from Ming Yuan et al. 2006

Sparse Structured FNN

- Pruning of Fully-connected Neural Networks

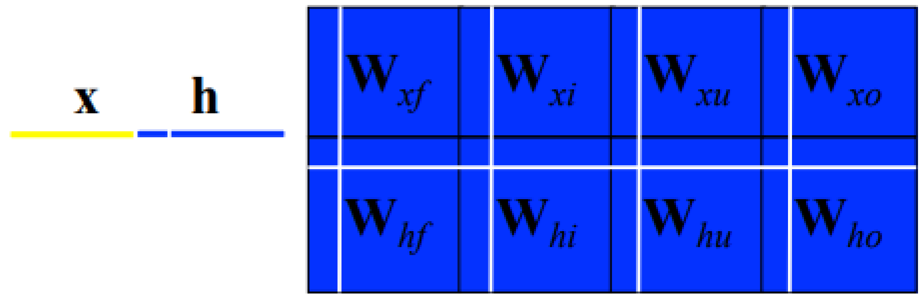




Sparse Structured LSTM

- Pruning of LSTMs

$$
\begin{aligned}
\boldsymbol{f}_t &= \sigma([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}]W_f + \boldsymbol{b}_f) \\
\boldsymbol{u}_t &= \tanh([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}]W_u + \boldsymbol{b}_c) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{u}_t \\
\boldsymbol{i}_t &= \sigma([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}]W_i + \boldsymbol{b}_i) \\
\boldsymbol{o}_t &= \sigma([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}]W_o + \boldsymbol{b}_o) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)
\end{aligned}
$$



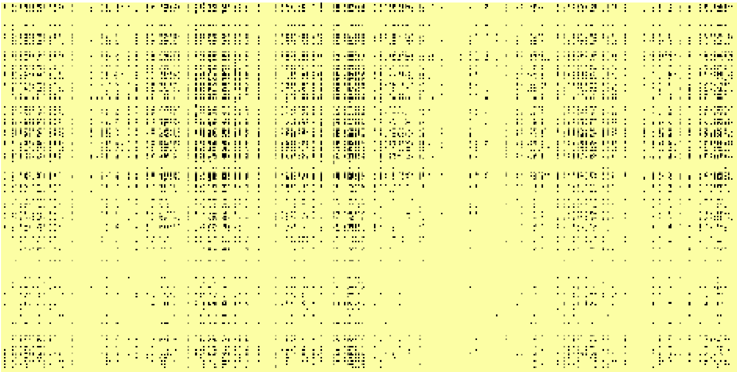
Weights matrices in LSTM

Weights in next layer(s)

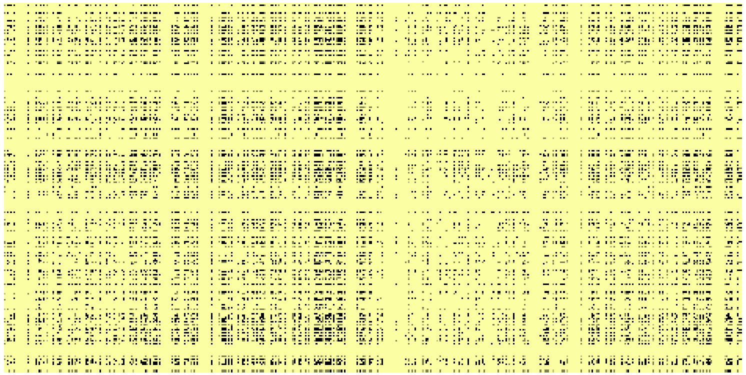
Figure from Wei Wen et al. 2017

Sparse Structured LSTM

LSTM layer 1



LSTM layer 2



	Structure	Params	FLOPs
Original model	1500-1500-1500	51M	100%
Pruned model	533-425-533	9M(18%)	18%

Toy Experiment on MNIST

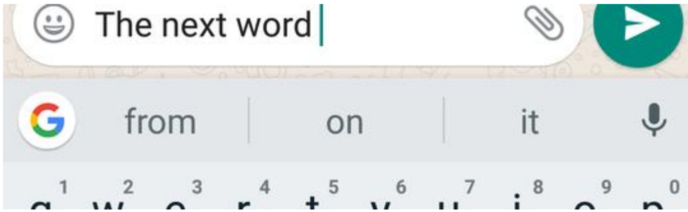
- Model: 784-300-100 fully-connected NN
- FLOPs for a matrix *W* is calculated as the size of the smallest sub-matrix formed by such rows and columns that contain all non-zero elements in *W.*
- GSP: group sparse prior
- PR: pruning and retraining

Method	Model	Params	FLOPs	Test Error (%)
SGD (baseline)	1 model	1*	1*	1.66
SGD	18 models	18×	18×	1.49
SGLD+GSP+PR	18 models†	1.8×	2.5×	**1.26**
SGLD+GSP+PR	18 models‡	**0.7**×	**2.2**×	1.29

* The baseline model has 266K parameters and 532K FLOPs.
† indicates 90% sparsity and ‡ indicates 96% sparsity for each model.

Language Modeling Experiment

- Language Modeling



- Penn Tree Bank dataset

  Vocabulary size: 10K

  Dataset size: 929K/73K/10K words in training, development and test sets respectively.

- 2-layers LSTM model

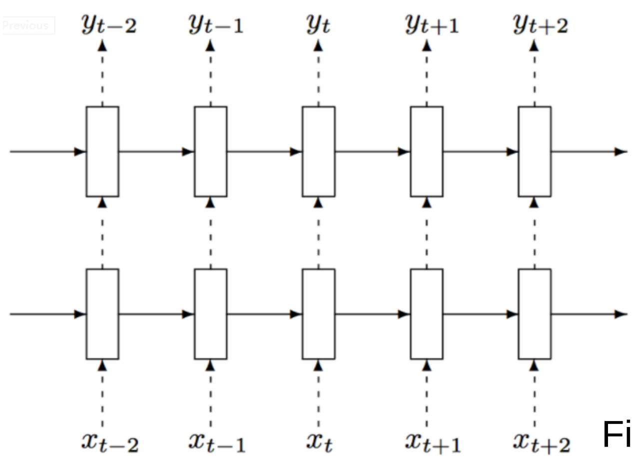


Figure from Zaremba et al. 2014

- Perplexity

  A measurement of how well the language model predicts the word sequence.

$$\mathrm{PPL} = e^{-\frac{1}{N}\sum \log P(w_i)}$$

Language Modeling Experiment

- Comparison of various models based on LSTMs on PTB dataset.

Method	Model	Params	FLOPs	Dev.	Test
SGD [10]	1 large	1*	1*	82.2	78.4
SGD [10]	38 large	38×	38×	71.9	68.7
VD [24]	10 large	10×	-	-	68.7
VD+SEAL [11]	individual	51M	-	71.1	68.5
SGLD+GSP+PR	20 large	2.0×	4.5×	68.6	66.4
SGLD+GSP+PR	4 large	0.4×	0.5×	72.2	69.7
SGLD+GSP+PR+SE	4 large	**0.3**×	**0.7** ×	**64.4**	**62.1**

* The baseline LSTM model has 66M parameters and 102M FLOPs.

Ref： [10] Wojciech Zaremba et al. 2014; [11] Hakan Inan et al. 2017; [24] Yarin Gal et al. 2016

Language Modeling Experiment

- Comparison of various models based on LSTMs on PTB dataset.

Method	Model	Params	FLOPs	Dev.	Test
SGD [10]	1 large	1*	1*	82.2	78.4
SGD [10]	38 large	38×	38×	71.9	68.7
VD [24]	10 large	10×	-	-	68.7
VD+SEAL [11]	individual	51M	-	71.1	68.5
SGLD+GSP+PR	20 large	2.0×	4.5×	68.6	66.4
SGLD+GSP+PR	4 large	0.4×	0.5×	72.2	69.7
SGLD+GSP+PR+SE	4 large	**0.3**×	**0.7** ×	**64.4**	**62.1**

* The baseline LSTM model has 66M parameters and 102M FLOPs.

Ref： [10] Wojciech Zaremba et al. 2014; [11] Hakan Inan et al. 2017; [24] Yarin Gal et al. 2016

Language Modeling Experiment

- Comparison of various models based on LSTMs on PTB dataset.

Method	Model	Params	FLOPs	Dev.	Test
SGD [10]	1 large	1*	1*	82.2	78.4
SGD [10]	38 large	38×	38×	71.9	68.7
VD [24]	10 large	10×	-	-	68.7
VD+SEAL [11]	individual	51M	-	71.1	68.5
SGLD+GSP+PR	20 large	2.0×	4.5×	68.6	66.4
SGLD+GSP+PR	4 large	0.4×	0.5×	72.2	69.7
SGLD+GSP+PR+SE	4 large	**0.3**×	**0.7** ×	**64.4**	**62.1**

* The baseline LSTM model has 66M parameters and 102M FLOPs.

Ref： [10] Wojciech Zaremba et al. 2014; [11] Hakan Inan et al. 2017; [24] Yarin Gal et al. 2016

Language Modeling Experiment

- Comparison of various models based on LSTMs on PTB dataset.

Method	Model	Params	FLOPs	Dev.	Test
SGD [10]	1 large	1^{*}	1^{*}	82.2	78.4
SGD [10]	38 large	$38\times$	$38\times$	71.9	68.7
VD [24]	10 large	$10\times$	-	-	68.7
VD+SEAL [11]	individual	51M	-	71.1	68.5
SGLD+GSP+PR	20 large	$2.0\times$	$4.5\times$	68.6	66.4
SGLD+GSP+PR	4 large	$0.4\times$	$0.5\times$	72.2	69.7
SGLD+GSP+PR+SE	4 large	**0.3**$\times$	**0.7** $\times$	**64.4**	**62.1**

* The baseline LSTM model has 66M parameters and 102M FLOPs.

Ref： [10] Wojciech Zaremba et al. 2014; [11] Hakan Inan et al. 2017; [24] Yarin Gal et al. 2016

Conclusion & Future Work

Conclusion:

- Propose a novel approach for learning ensembles of neural networks.
- Combination of SG-MCMC sampling, group sparse prior and network pruning.
- Experimental verifications for sparse structure learning for LSTM models.

Future work:

- Interleaving model sampling and model pruning.
- Expand to more tasks.

Thank you!

Speaker: Yichi Zhang
E-mail: zhangyic17@mails.tsinghua.edu.cn