Improving Neural Language Modeling via Adversarial Training

Dilin Wang*, Chengyue Gong* (equal contribution) Qiang Liu

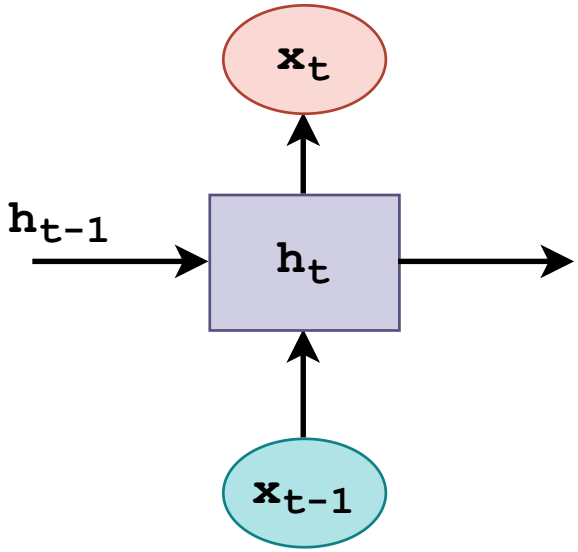
Department of Computer Science
The University of Texas at Austin

Neural Language Modeling

- Example: `the clouds are in the` sky

$$h_t = f_{NN}(x_{t-1}, h_{1:t-1}; \boldsymbol{\theta})$$

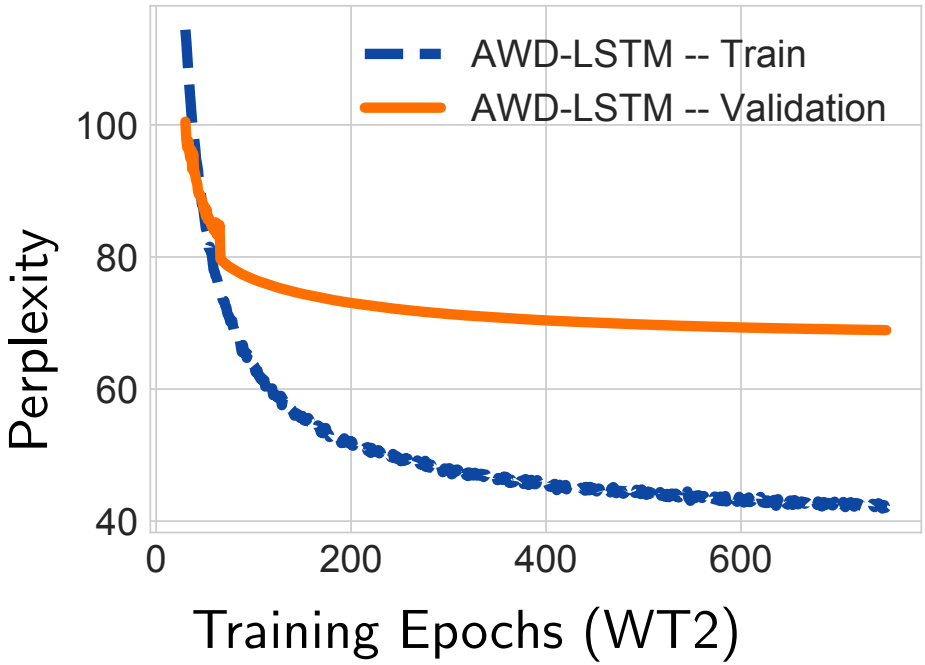
$$p(x_t \mid x_{1:t-1}; \boldsymbol{\theta}, \boldsymbol{w}) = \mathrm{Softmax}(x_t, h_t; \boldsymbol{w}) = \frac{\exp(w_{x_t}^\top h_t)}{\sum_{\ell=1}^{|\mathcal{V}|} \exp(w_\ell^\top h_t)}$$



- Maximum log-likelihood estimation (MLE):

$$\max_{\boldsymbol{\theta}, \boldsymbol{w}} \quad \sum_t \log p(x_t | x_{1:t-1}; \boldsymbol{\theta}, \boldsymbol{w})$$

Overfitting



- Existing overfitting preventing methods:
  - Dropout[e.g., Gal & Ghahramani, 2016]
  - Optimizer [e.g., Merity et al., 2017]
  - Other: weight tying [Press & Wolf, 2016; Inan et al., 2017]; activation regularization [Merity et al., 2017], etc.

Adversarial MLE

- Idea: inject an adversarial perturbation on the word embedding vectors in the Softmax layer, and maximize the worst-case performance,

$$\max_{\boldsymbol{\theta},\boldsymbol{w}} \min_{\delta_t} \sum_t \log \left(\frac{\exp((w_t + \delta_t)^\top h_t)}{\exp((w_t + \delta_t)^\top h_t) + \sum_{j \neq t} \exp(w_j^\top h_t)} \right)$$

$$s.t \quad ||\delta_t|| \leq \epsilon.$$

A closed-form solution

$$\delta_t^* = \arg\min_{||\delta_t|| \leq \epsilon} (w_t + \delta_t)^\top h_t = -\epsilon \frac{h_t}{||h_t||}.$$

Adversarial MLE

- Idea: inject an adversarial perturbation on the word embedding vectors in the Softmax layer, and maximize the worst-case performance,

$$\max_{\boldsymbol{\theta},\boldsymbol{w}} \min_{\delta_t} \sum_t \log \left(\frac{\exp((w_t + \delta_t)^\top h_t)}{\exp((w_t + \delta_t)^\top h_t) + \sum_{j \neq t} \exp(w_j^\top h_t)} \right)$$

$$s.t \quad ||\delta_t|| \leq \epsilon.$$

A closed-form solution

$$\delta_t^* = \arg\min_{||\delta_t|| \leq \epsilon} (w_t + \delta_t)^\top h_t = -\epsilon \frac{h_t}{||h_t||}.$$

Adversarial MLE Promotes Diversity

- If w_i dominates all the other words under ϵ-adversarial perturbation, in that

$$\min_{||\delta_i|| \leq \epsilon} (w_i + \delta_i)^\top h = (w_i^\top h - \epsilon ||h||)$$
$$> w_j^\top h, \quad \forall j \neq i,$$

then we have,

$$\min_{j \neq i} ||w_j - w_i|| > \epsilon,$$

that is, w_i is separated from the embedding vectors of all other words by at least ϵ distance.

Improving on Language Modeling

Method	Params	Valid	Test
AWD-LSTM (Merity et al., 2017)	24M	51.60	51.10
AWD-LSTM + Ours	24M	**49.31**	**48.72**
AWD-LSTM + MoS (Yang et al., 2017)	22M	48.33	47.69
AWD-LSTM + MoS + Ours	22M	**47.15**	**46.52**

Table: PTB

Method	Params	Valid	Test
AWD-LSTM (Merity et al., 2017)	33M	46.40	44.30
AWD-LSTM + Ours	33M	**42.48**	**40.71**
AWD-LSTM + MoS (Yang et al., 2017)	35M	42.41	40.68
AWD-LSTM + MoS + Ours	35M	**40.27**	**38.65**

Table: WT2

Improving on Machine Translation

Method	BLEU
Transformer Base Vaswani et al., 2017	27.30
Transformer Base + Ours	**28.43**
Transformer Big Vaswani et al., 2017	28.40
Transformer Big + Ours	**29.52**

Table: WMT2014 Ee→De

Method	BLEU
Transformer Small Vaswani et al., 2017	32.47
Transformer Small + Ours	**33.61**
Transformer Base Wang et al., 2018	34.43
Transformer Base + Ours	**35.18**

Table: IWSLT2014 De→En

Conclusions

Proposed an adversarial training mechanism for language modeling

1. A Closed-form solution & easy to implement
2. Diversity Promotion
3. Strong empirical results

Thank You

Poster #105, Today 06:30 PM – 09:00 PM @ Pacific Ballroom