Data-dependent Gaussain Prior Objective for Language Generation

2020 NLP Study

International Conference on Learning Representations (ICLR) 2020

Presented by Miyoung Ko

Background

Language Generation

- Langauge model, NMT, Text summarization, Image captioning etc...
- Sequence predictions

$$\boldsymbol{y} \sim p_\theta(\boldsymbol{x}) \,, \quad \boldsymbol{y} = \langle y_1, y_2, ..., y_l \rangle$$

Probability : $p_\theta(\boldsymbol{y}|\boldsymbol{x}) = p_\theta(y_1|\boldsymbol{x})p_\theta(y_2|\boldsymbol{x}, y_1)...p_\theta(y_l|\boldsymbol{x}, y_{1:l-1})$

Maximum likelihood estimation (MLE) Loss:

$$\mathcal{L}_{\mathbf{MLE}}(\theta) = -\log p_\theta(\boldsymbol{y}|\boldsymbol{x}) = -\sum_{i=1}^{l} \log p_\theta(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i}).$$

Background

Maximum likelihood estimation (MLE) Loss:

$$\mathcal{L}_{\mathbf{MLE}}(\theta) = -\log p_\theta(\boldsymbol{y}|\boldsymbol{x}) = -\sum_{i=1}^{l} \log p_\theta(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i}).$$

Exposure Bias : Models is not exposed to the full range of erros during training

➔ Reinforcement learning models (training sequences are generated by the model itself)

Loss Mismatch : Maximize log-likelihood during training, evaluate on different metric (BLUE, ROUGE)

➔ MIXER, minimum dibergence, mzximum margin

Generation Diversity : Generations are dull, generic, repetitive, short-sighted

➔ Adding linguistic features to latent variable and etc…

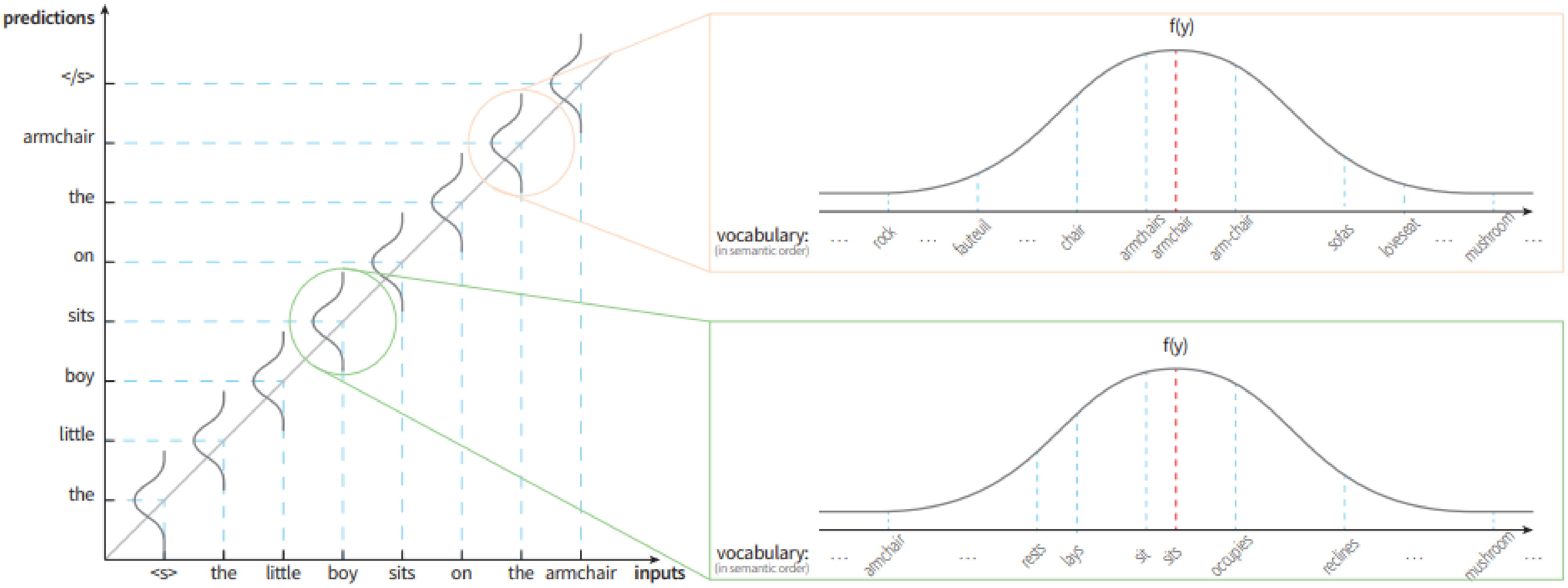
Negative Dibersity Ignorance : MLE can't assign scores to different incorrect model outputs, All incorrect outputs are treated equally during training

Armchair ➔ *Deckchair vs. Mushroom*

D2GPo

D2GPo : Data-dependent Gaussian Prior Objective

- Add extra Gaussian prior objective to MLE loss (KL divergence loss term)
- KL divergence b/w model training prediction and data-dependent Gaussian prior distribution



D2GPo

D2GPo : Data-dependent Gaussian Prior Objective

- Add extra Gaussian prior objective to MLE loss (KL divergence loss term)
- KL divergence b/w model training prediction and data-dependent Gaussian prior distribution

Evaluation Function : $f(\tilde{\boldsymbol{y}}, \boldsymbol{y}) \in \mathbb{R}$ ➔ Higher value indicates a better probaility

$$\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q) = KL(q(\boldsymbol{y}) \| p_\theta(\boldsymbol{y}|\boldsymbol{x})) - \alpha \mathbb{E}_q \left[f(\tilde{\boldsymbol{y}}, \boldsymbol{y}) \right],$$

Prior distribution q from the ground-truth data (independent of model parameters) ➔ $\mathbb{E}_q \left[f(\tilde{\boldsymbol{y}}, \boldsymbol{y}) \right] = 0.$

$$\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q) = KL(q(\boldsymbol{y}) \| p_\theta(\boldsymbol{y}|\boldsymbol{x})),$$

$$KL(q \| p_\theta) = \mathbb{E}_p(\log(\frac{q}{p})) = \sum_i q_i * \log(q_i) - \sum_i q_i * \log(p_i).$$

$$\min_\theta \mathcal{L}_{\mathbf{MLE}}(\boldsymbol{\theta}) + \lambda \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q),$$

D2GPo

D2GPo : Data-dependent Gaussian Prior Objective

$$\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q) = KL(q(\boldsymbol{y}) \| p_{\theta}(\boldsymbol{y}|\boldsymbol{x})),$$

$$KL(q\|p_\theta) = \mathbb{E}_p(\log(\frac{q}{p})) = \sum_i q_i * \log(q_i) - \sum_i q_i * \log(p_i).$$

$$q(y^*) = \frac{exp(f(\tilde{\boldsymbol{y}}, y^*)/T)}{\sum_j exp(f(\tilde{y}_j, y^*)/T)},$$

Evaluation Function :

$$dist_{i,j} = cosine\_similarity(emb(y_i), emb(\tilde{y}_j)).$$

$$ORDER(y_i) = sort([dist_{i,1}, dist_{i,2}, ..., dist_{i,N}]).$$

Discussion:

- Evaluation function f of q -> Gaussian probaibility density function
- Linear additive property of word embedding (king – man + woman = queen)

➔ Gaussian distribution for the embedding-distance-determined order

- Different from data-independent Gaussian prior like L2 regularization (zero-mean Gaussian)

Experiments and Results

Embedding Pre-training

- Either word embeddings or byte pair-encoding (BPE)
- fastText 512 dim, window size 5, 10 negative samples
- NMT : cross-ligual BPE subword embedding
- Text summarization and others : BPE subword embedding on English monolingual corpora

1) Supervised NMT

- WMT 14 EN-DE / EN-FR, WMT 16 EN-RO

System	EN–DE	EN–FR	EN–RO	EN–RO + STD
Vaswani et al. (2017) (base)	27.30	38.10	-	-
Vaswani et al. (2017) (big)	28.40	41.00	-	-
Transformer (base)	27.35	38.44	33.22	36.68
+ **D2GPo**	**27.93++**	**39.23++**	**34.00+**	**37.11+**
Transformer (big)	28.51	41.05	33.45	37.55
+ **D2GPo**	**29.10+**	**41.77++**	**34.13+**	**37.92+**

Experiments and Results

2) Unsupervised NMT

Method	EN–FR	FR–EN	EN–DE	DE–EN	EN–RO	RO–EN
Artetxe et al. (2017)	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	33.40	33.30	27.00	34.30	33.30	31.80
MASS (Song et al., 2019)	37.50	34.90	28.30	35.20	35.20	33.10
MASS + **D2GPo**	**37.92**	**34.94**	**28.42**	**35.62**	**36.31**	**33.41**

3) Text Summarization

	Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Supervised	RNN-based seq2seq	35.50	15.54	32.45
	Nallapati et al. (2016)	34.97	17.17	32.70
Semi-supervised	MLM pre-training (Song et al., 2019)	37.75	18.45	34.85
	DAE pre-training (Song et al., 2019)	35.97	17.17	33.14
	MASS pre-training (Song et al., 2019)	38.73	19.71	35.96
	MASS + **D2GPo**	**39.23**	**20.11**	**36.48**

Experiments and Results

4) Storytelling

Model	Params	Valid Perplexity	Test Perplexity
GCNN LM	123.4 M	54.50	54.79
GCNN + self-attention LM	126.4 M	51.84	51.18
LSTM seq2seq	110.3 M	46.83	46.79
Conv seq2seq	113.0 M	45.27	45.54
Conv seq2seq + self-attention	134.7 M	37.37	37.94
Ensemble: Conv seq2seq + self-attention	270.3 M	36.63	36.93
Fusion: Conv seq2seq + self-attention	255.4 M	36.08	36.56
Conv seq2seq + self-attention + **D2GPo**	134.7 M	**35.56**	**35.74**
Fusion: Conv seq2seq + self-attention + **D2GPo**	255.4 M	**33.82**	**33.90**

5) Image Captioning

	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Att2in (Rennie et al., 2017)	-	31.3	26.0	54.3	101.3	-
Att2all (Rennie et al., 2017)	-	30.0	25.9	53.4	99.4	-
Baseline: Top-down	74.5	33.4	26.1	54.4	105.4	19.2
Baseline + **D2GPo**	**75.2**	**33.6**	**26.3**	**55.1**	**106.6**	**19.7**
Baseline + SCST	77.8	34.4	26.6	56.1	114.3	19.9
Baseline + SCST + **D2GPo**	**78.0**	**34.7**	**26.8**	**56.3**	**116.8**	**20.2**

Experiments and Results

4) Storytelling

Model	Params	Valid Perplexity	Test Perplexity
GCNN LM	123.4 M	54.50	54.79
GCNN + self-attention LM	126.4 M	51.84	51.18
LSTM seq2seq	110.3 M	46.83	46.79
Conv seq2seq	113.0 M	45.27	45.54
Conv seq2seq + self-attention	134.7 M	37.37	37.94
Ensemble: Conv seq2seq + self-attention	270.3 M	36.63	36.93
Fusion: Conv seq2seq + self-attention	255.4 M	36.08	36.56
Conv seq2seq + self-attention + **D2GPo**	134.7 M	**35.56**	**35.74**
Fusion: Conv seq2seq + self-attention + **D2GPo**	255.4 M	**33.82**	**33.90**

5) Image Captioning

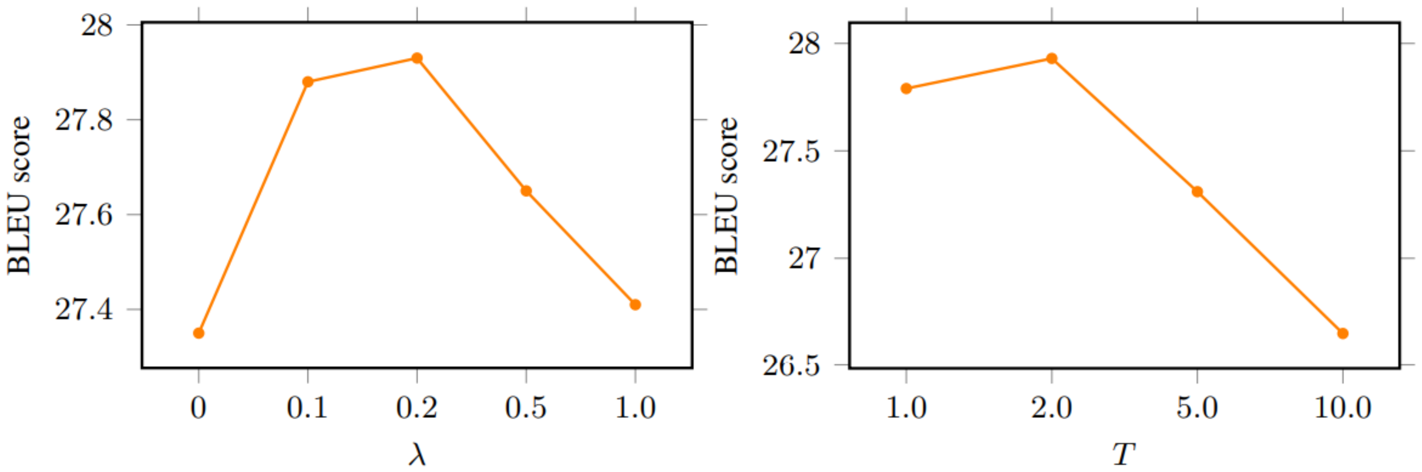
	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Att2in (Rennie et al., 2017)	-	31.3	26.0	54.3	101.3	-
Att2all (Rennie et al., 2017)	-	30.0	25.9	53.4	99.4	-
Baseline: Top-down	74.5	33.4	26.1	54.4	105.4	19.2
Baseline + **D2GPo**	**75.2**	**33.6**	**26.3**	**55.1**	**106.6**	**19.7**
Baseline + SCST	77.8	34.4	26.6	56.1	114.3	19.9
Baseline + SCST + **D2GPo**	**78.0**	**34.7**	**26.8**	**56.3**	**116.8**	**20.2**

Experiments and Results

Evaluation Function

Evaluation Function	BLEU	△
Baseline	27.35	
Gaussian	27.93	0.58 ↑
Random	26.34	1.01 ↓
Linear	27.45	0.10 ↑
Cosine	27.62	0.27 ↑

Hyperparameters



Low-resource setting

Method	10K	100K	600K
Baseline	1.01	17.80	33.22
+ **D2GPo**	4.33	20.48	34.00

Generation diversity

	#GOLD	Baseline	**+D2GPo**
#LF	4915	3900	3998
#SUM	63086	55234	56129
#RATIO	7.79%	7.06%	7.12%