On the Discrepancy between Density Estimation and Sequence Generation

Olga Bonetskaya

Is Log-likelihood correlated to BLEU?

Task: sequence-to-sequence generation

Metrics: test log-likelihood, BLEU

$$L(F) = \frac{1}{N} \sum_{n=1}^{N} \log p_F(y_n | x_n).$$

Models: autoregressive and latent variable models

Log-likelihood is highly correlated with BLEU when considering models within the same family. Log-likelihood is not correlated with BLEU when comparing models from different families.

Data

Datasets:

- IWSLT’16 De→En (197K training, 2K development, 2K test)
- WMT’16 En↔Ro (612K training, 2K development, 2K test)
- WMT’14 En↔De 4 (4.5M training, 3K development, 3K test)

Wordpiece tokenization

Knowledge distillation (Transformer-base, Transformer-small)

Autoregressive Models

Learning: $$\log p_{\mathrm{AR}}(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^{T} \log p_\theta(y_t|y_{<t}, \mathbf{x}).$$

$$L_{\mathrm{AR}}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log p_{\mathrm{AR}}(\mathbf{y}_n|\mathbf{x}_n).$$

Inference: $$\mathrm{argmax}_{\mathbf{y}} \log p_\theta(\mathbf{y}|\mathbf{x}) = \mathrm{argmax}_{y_{1:T}} \sum_{t=1}^{T} \log p_\theta(y_t|y_{<t}, \mathbf{x}).$$

Autoregressive Models

- Transformer-big (Tr-L)
- Transformer-base (Tr-B)
- Transformer-small (Tr-S)

Beam search is used

Latent Variable Models

Learning: $\log p_{\mathrm{LVM}}(\mathbf{y}|\mathbf{x}) = \log \int_{\mathbf{z}} p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x})\, p_\theta(\mathbf{z}|\mathbf{x}) d\mathbf{z}.$

$$\log p_{\mathrm{LVM}}(\mathbf{y}|\mathbf{x}) \geq \mathrm{ELBO}(\mathbf{y}, \mathbf{x}; \theta, \phi) \; = \mathop{\mathbb{E}}_{\mathbf{z}\sim q_\phi}\Big[\log p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x})\Big] - \mathrm{KL}\Big[q_\phi(\mathbf{z}|\mathbf{y}, \mathbf{x}) \,\Big\|\, p_\theta(\mathbf{z}|\mathbf{x})\Big]$$

Inference:

$$\delta(\mathbf{z}|\boldsymbol{\mu}) = \begin{cases} 1, & \text{if } \mathbf{z} = \boldsymbol{\mu} \\ 0, & \text{otherwise} \end{cases}$$

Then, the ELBO reduces to: $\log p_\theta(\mathbf{y}|\boldsymbol{\mu}, \mathbf{x}) + \log p_\theta(\boldsymbol{\mu}|\mathbf{x}).$

Factorization: $p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x}) \; = \; \prod_{t=1}^{T} p_\theta(y_t|\mathbf{z}, \mathbf{x}).$

Diagonal Gaussian vs Normalizing Flow

$$\log p_\theta(z_{1:T}|\mathbf{x}) = \sum_{t=1}^{T} \log \mathcal{N}\Big(z_t \Big| \mu_{\theta,t}(\mathbf{x}), \sigma_{\theta,t}(\mathbf{x})\Big)$$

a base distribution $p_b(\epsilon)$

$$f(\mathbf{z}) = \epsilon,\ f^{-1}(\epsilon) = \mathbf{z}$$

$$f(\mathbf{z};\mathbf{x}) = \epsilon,\ \ f^{-1}(\epsilon;\mathbf{x}) = \mathbf{z}.$$

$$\log p_\theta(\mathbf{z}|\mathbf{x}) = \log p_b\Big(f(\mathbf{z};\mathbf{x})\Big) + \log\left|\det \frac{\partial f(\mathbf{z};\mathbf{x})}{\partial \mathbf{z}}\right|$$

$$\begin{aligned}
\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{tr}} &= \text{split}(\mathbf{z}) \\
\mathbf{s}, \mathbf{b} &= g_{\text{param}}(\mathbf{z}_{\text{id}}) \\
f(\mathbf{z}) &= \text{concat}(\mathbf{z}_{\text{id}};\ \ \mathbf{s} \cdot \mathbf{z}_{\text{tr}} + \mathbf{b})
\end{aligned}$$

Latent Variable Models

encoder — Transformer encoder

length predictor — a 2-layer MLP

prior:

- Gauss (Transformer, a sequence of positional encodings of length T as input, outputs the mean and standard deviation)
- Normalizing Flow (Transformer decoder + Linear with weight-normalization,)

decoder — Transformer decoder

posterior — final Linear layer with weight normalization

Test BLEU score and log-likelihood of each model

		BLEU (↑)		LL (↑)	
		RAW	DIST.	RAW	DIST.
WMT'14 EN→DE	TR-S	24.54	24.94	-1.77	-2.36
	TR-B	28.18	27.86	-1.44	-2.19
	TR-L	<u>29.39</u>	28.29	-1.35	-2.23
	GA-B	15.74	24.54	-1.51	-2.44
	GA-L	17.33	**25.53**	-1.47	-2.24
	FL-S	18.17	21.98	-1.41	-2.13
	FL-B	18.57	21.82	**-1.23**	-2.05
	FL-B$^{(*)}$	18.55	21.45		
	FL-L$^{(*)}$	20.85	23.72		
WMT'14 DE→EN	TR-S	29.15	28.40	-1.66	-2.24
	TR-B	32.21	32.24	-1.42	-2.12
	TR-L	<u>33.16</u>	32.24	-1.35	-2.05
	GA-B	21.64	29.29	-1.41	-2.17
	GA-L	23.03	**30.30**	-1.31	-2.04
	FL-S	23.17	27.14	-1.28	-1.73
	FL-B	23.12	26.72	**-1.20**	-1.71
	FL-B$^{(*)}$	23.36	26.16		
	FL-L$^{(*)}$	25.40	28.39		

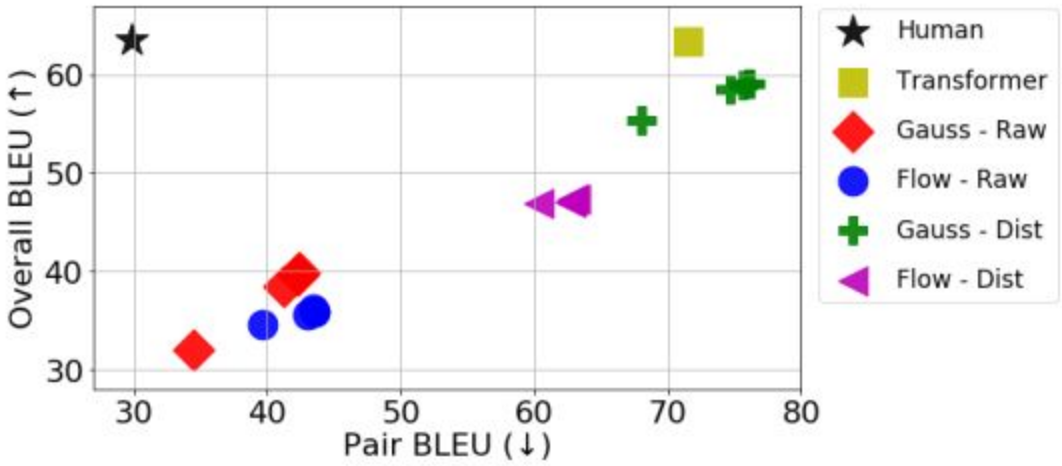
Pearson's correlation between log-likelihood and BLEU

	TR-B	GA-B	FL-B
RAW	0.926	0.831	0.678
DIST.	-0.758	-0.897	-0.873

BLEU scores and log-likelihoods on out-of-distribution test sets.

		BLEU (↑)		LL (↑)	
		RAW	DIST.	RAW	DIST.
WMT'14 →IWSLT	TR-S	29.15	28.40	-1.65	-2.25
	TR-B	32.29	31.75	-1.42	-2.12
	TR-L	<u>33.16</u>	32.24	-1.35	-2.06
	GA-B	24.26	28.77	-1.37	-2.10
	GA-L	25.46	**29.60**	-1.28	-2.01
	FL-S	24.35	26.79	-1.26	-1.76
	FL-B	24.25	27.12	**-1.19**	-1.73
IWSLT→ WMT'14	TR-S	18.50	<u>18.94</u>	-2.79	-3.41
	GA-B	12.12	13.78	-3.10	-3.83
	FL-S	11.78	**14.35**	-2.81	-3.22
	FL-B	12.56	14.30	**-2.62**	-3.43

Results



Overall BLEU (↑)
Pair BLEU (↓)
30
40
50
60
70
80
Human
Transformer
Gauss - Raw
Flow - Raw
Gauss - Dist
Flow - Dist

Results

	BLEU					SPEED					SIZE
$k =$	0	1	2	4	8	0	1	2	4	8	
TR-S	24.54					2.69					17M
TR-B	28.18					2.58					60M
TR-L	29.39					1.93					208M
GA-B	23.15	24.54	24.87	24.94	24.92	28.77	20.52	16.51	12.00	8.11	75M
GA-L	24.31	25.53	25.69	25.68	25.68	19.83	14.72	10.25	7.88	4.91	95M
FL-B	21.57	21.82	21.79	21.81	21.80	5.82	5.60	4.84	3.60	3.37	75M
FL-L$^{(*)}$	23.72										258M

Вопросы

1) Перед вами таблица с результатами эксперимента для трёх моделей на одних и тех же данных. Предположите, из каких семейств (из одного или из разных) эти модели, и объясните, почему вы так думаете.

BLEU (↑)		LL (↑)	
RAW	DIST.	RAW	DIST.
24.54	24.94	-1.77	-2.36
28.18	27.86	-1.44	-2.19
<u>29.39</u>	28.29	-1.35	-2.23

2) Расскажите про корреляцию между BLEU и LL для моделей внутри одного семейства и из разных семейств.

3) Опишите модели LVM, которые использовались для эксперимента.