On the Learning of Non-Autoregressive Transformers

Fei Huang*, Tianhua Tao*, Hao Zhou, Lei Li, Minlie Huang

*: Equal Contribution

CoAI group, Tsinghua University

ByteDance AI Lab

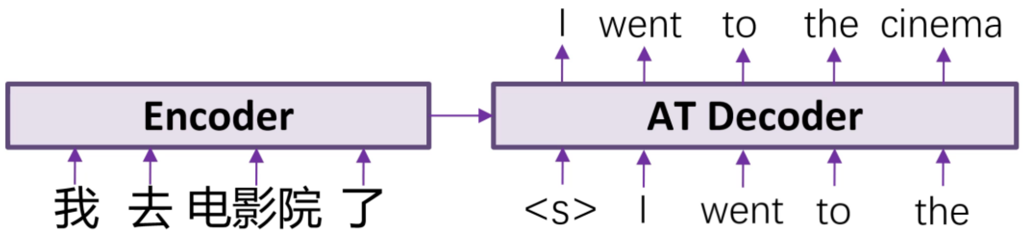
Institute for AI Industry Research, Tsinghua University

University of California, Santa Barbara

Background

Autoregressive Transformer (AT)

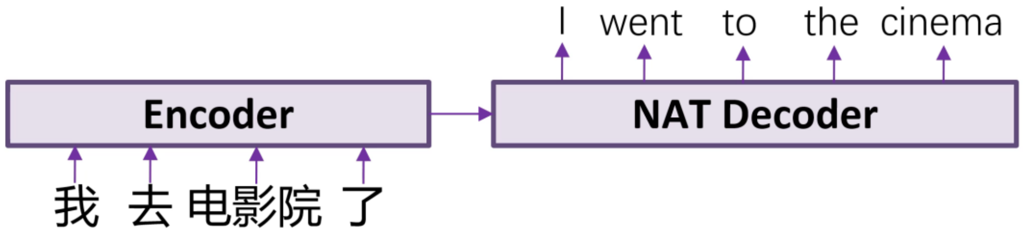
$$\log P_\theta^{\text{AT}}(Y|X) = \sum_{i=1}^{M} \log P_\theta^{\text{AT}}(y_i|y_{<i}, X)$$



Reduce the inference latency

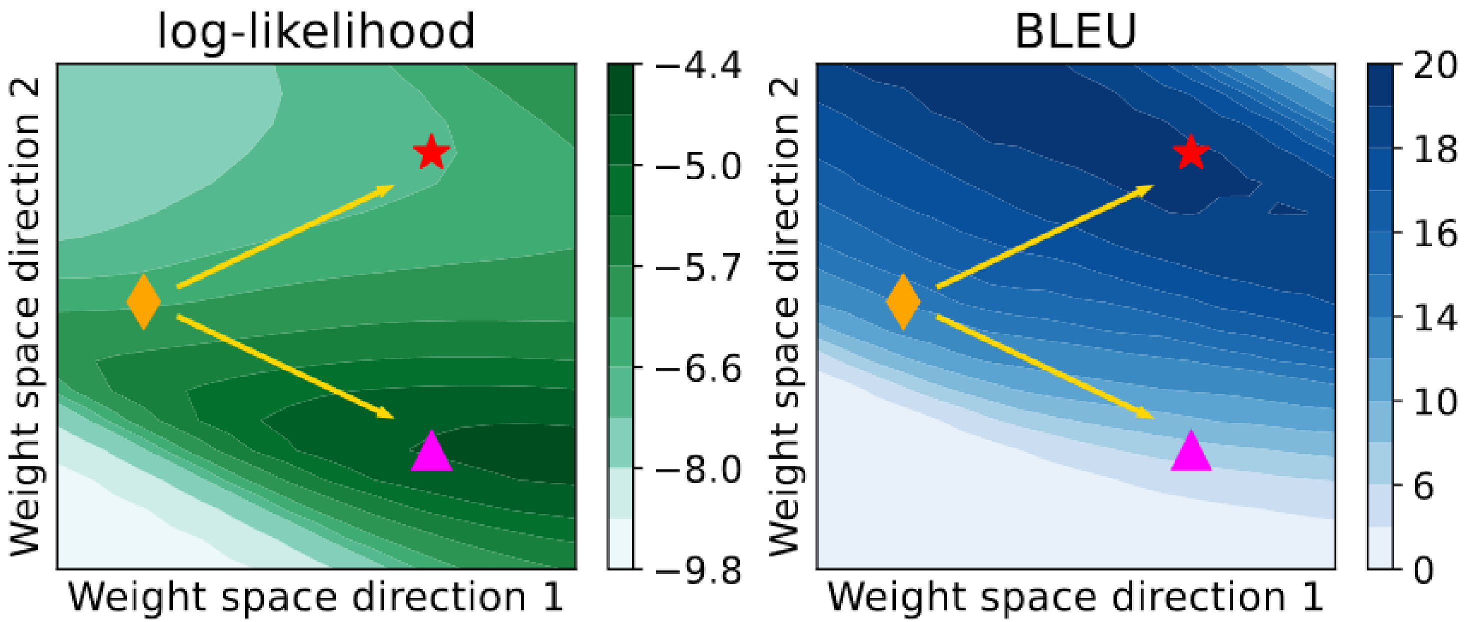
Non-Autoregressive Transformer (NAT)

$$\log P_\theta^{\text{NAT}}(Y|X) = \sum_{i=1}^{M} \log P_\theta^{\text{NAT}}(y_i|X)$$



NAT Learning is Challenging

- Maximum Likelihood Estimation (MLE) does not leads to higher BLEU



Start Checkpoint (◆)   MLE (▲)   GLAT+KD (★)

GLAT: Qian et al. Glancing Transformer for Non-Autoregressive Neural Machine Translation. ACL2021

KD: Kim and Rush. Sequence-level Knowledge Distillation. EMNLP2016

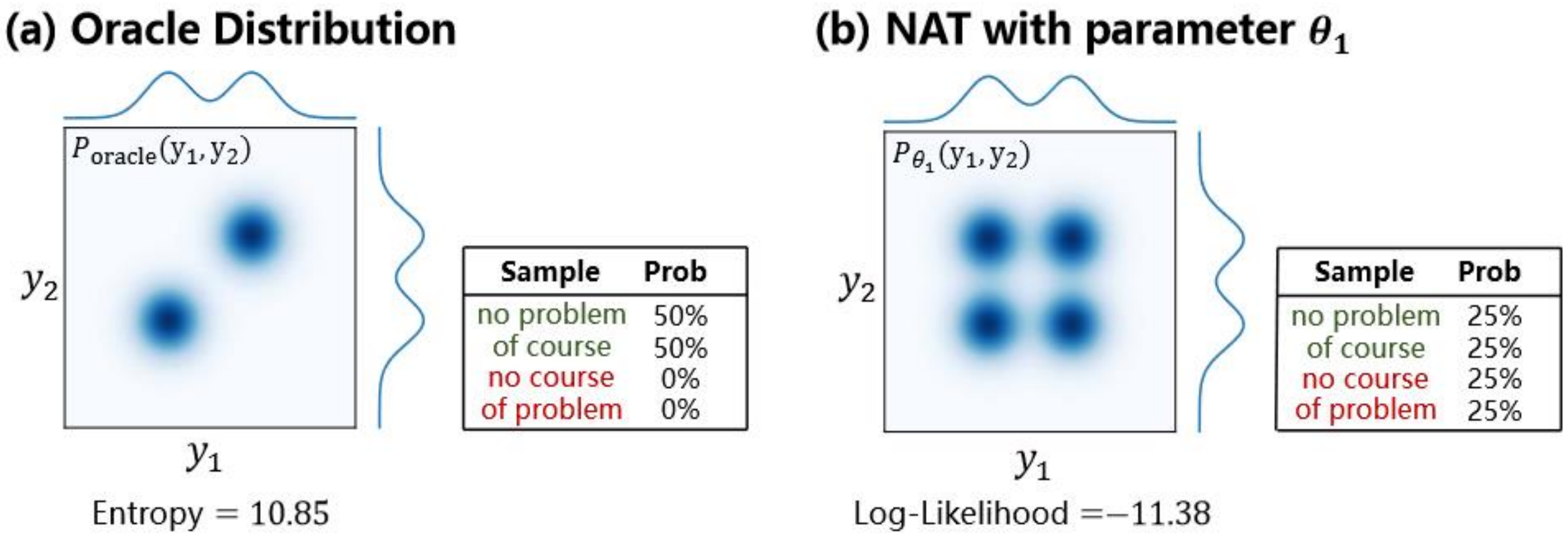
Two Questions

Q1: Why is NAT learning **so challenging** that **MLE does not work well**?

Q2: Why are **previously proposed objectives successful** despite they lead to **low likelihood**?

Q1: Why MLE does not work well

- **From intuitive perspective**



  - Maximizing the likelihood leads to **an approximation of marginal distributions**, but **drops the dependencies between tokens**

Q1: Why MLE does not work well

- **From theoretical perspective**

Theorem 1. *For a NAT model* $P_\theta(Y|X)$, *we have* $\min_\theta \mathcal{D}_{KL}[P_{data}(Y|X)||P_\theta(Y|X)] \geq \mathcal{C}$.

$$\mathcal{C} := \sum_{i=1}^{M} H_{\text{data}}(y_i|X) - H_{\text{data}}(Y|X)$$

C is a **property** of $P_{data}(Y|X)$, called ***Conditional Total Correlation*** (Watanabe, 1960)
It measures **the information of dependencies between target tokens**.

Watanabe. Information theoretical analysis of multi-variate correlation. *IBM J. Res. Dev.*, 1960.

Q1: Why MLE does not work well

- **From theoretical perspective**

Theorem 1. *For a NAT model $P_\theta(Y|X)$, we have* $\min_\theta \mathcal{D}_{KL}[P_{data}(Y|X)||P_\theta(Y|X)] \geq \mathcal{C}$.

$$\mathcal{C} := \sum_{i=1}^{M} H_{\text{data}}(y_i|X) - H_{\text{data}}(Y|X)$$

$\mathcal{C}$ is a **property** of $P_{data}(Y|X)$, called ***Conditional Total Correlation*** (Watanabe, 1960)
It measures **the information of dependencies between target tokens**.

Remark 1. *For a well-trained NAT in terms of KL divergence, **the dropped information can be measured by** $\mathcal{C}$.*

Watanabe. Information theoretical analysis of multi-variate correlation. *IBM J. Res. Dev.*, 1960.

Q1: Why MLE does not work well

- **From theoretical perspective**

Theorem 1. *For a NAT model $P_\theta(Y|X)$, we have* $\min_\theta \mathcal{D}_{KL}[P_{data}(Y|X)||P_\theta(Y|X)] \geq \mathcal{C}$.

$$\mathcal{C} := \sum_{i=1}^{M} H_{\text{data}}(y_i|X) - H_{\text{data}}(Y|X)$$

$\mathcal{C}$ is a **property** of $P_{data}(Y|X)$, called ***Conditional Total Correlation*** (Watanabe, 1960)
It measures **the information of dependencies between target tokens**.

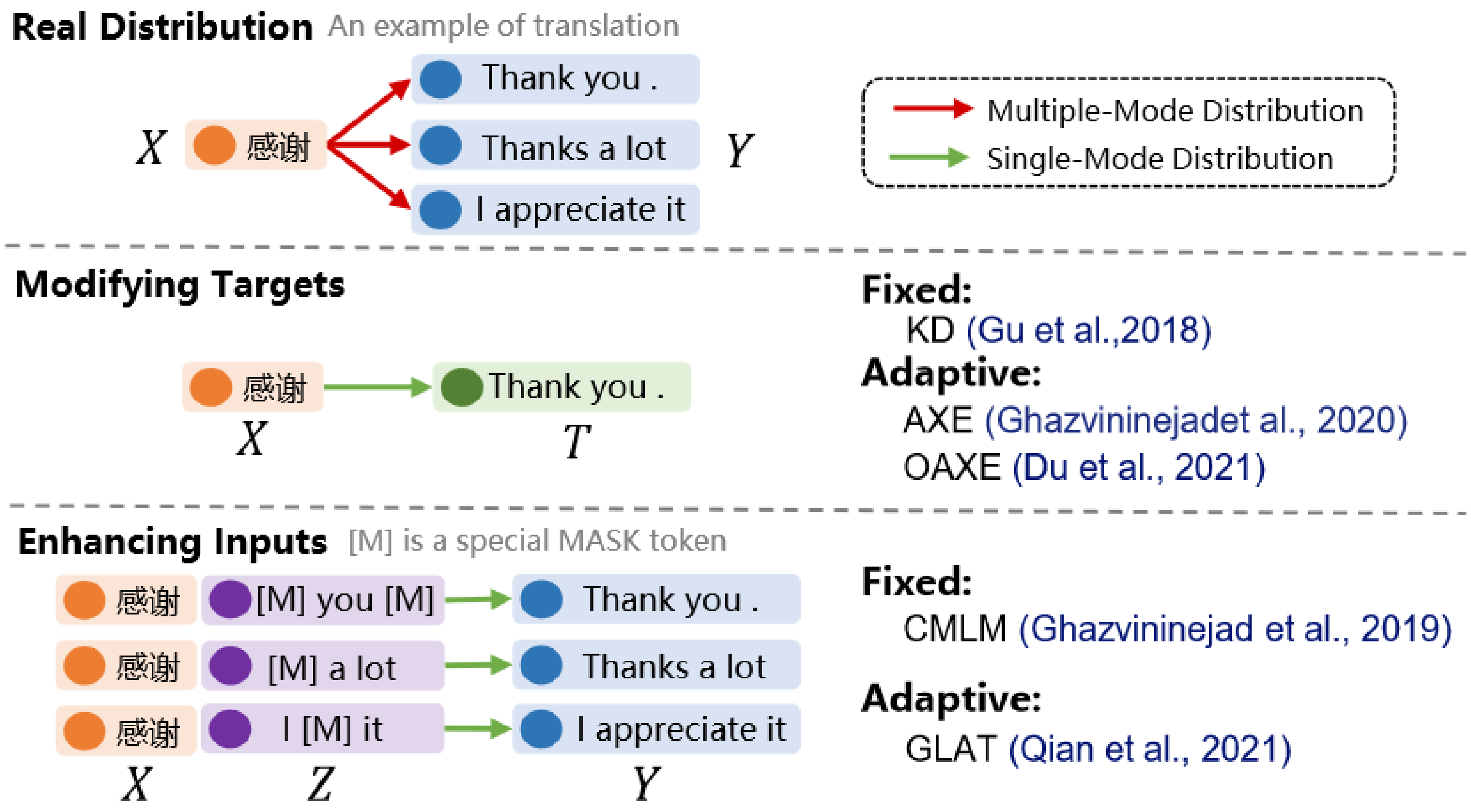
Remark 1. *For a well-trained NAT in terms of KL divergence,* ***the dropped information can be measured by*** $\mathcal{C}$.

Remark 2. $\mathcal{C}$ ***represents the difficulties of NAT learning***. *Given the data distribution, an NAT cannot achieve an information loss less than* $\mathcal{C}$*, regardless of its parameters or training methods.*

Watanabe. Information theoretical analysis of multi-variate correlation. *IBM J. Res. Dev.*, 1960.

Q2: Why previous objectives successful?

- Revisit previous objectives



Q2: Why previous objectives successful?

- **A Unified Perspective – Maximum Proxy-Likelihood Estimation**
  - Proxy Distribution $Q(T|X,Z)$

Q2: Why previous objectives successful?

- **A Unified Perspective – Maximum Proxy-Likelihood Estimation**
  - Proxy Distribution $Q(T|X,Z)$
  - Unified Objective

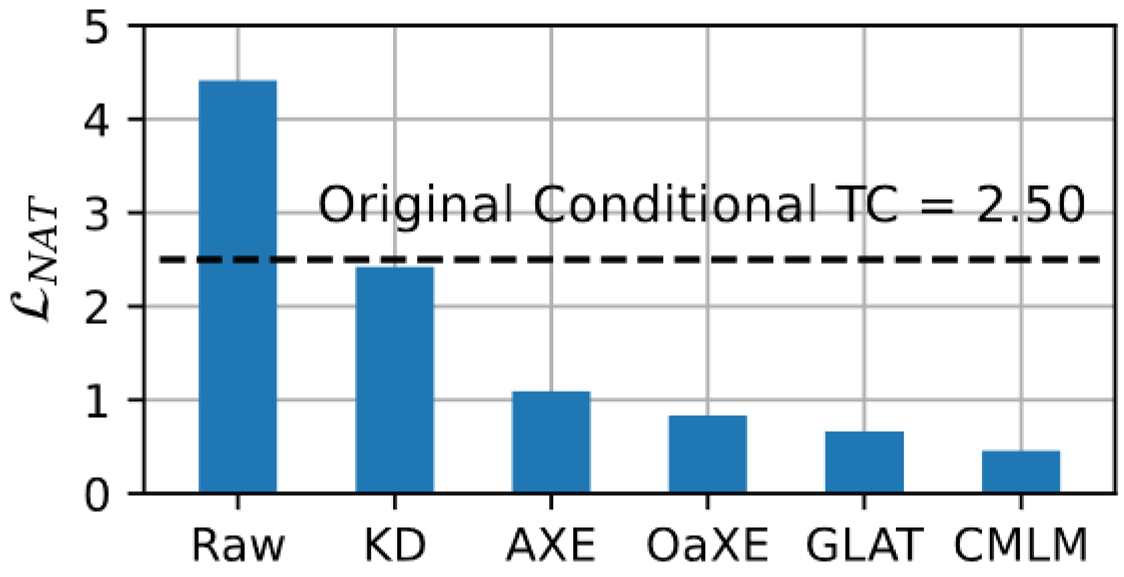
$$\mathcal{L}_{\mathbf{MPLE}} = \underbrace{\mathcal{D}_{\mathrm{KL}}(Q||P_\theta)}_{\text{MLE objective on } Q \text{ instead of } P_{data}} + \underbrace{\mathcal{R}(Q, P_{\mathrm{data}})}_{\text{Data distortion between } Q \text{ and } P_{data}}$$

Q can have a smaller C than P_{data}, thereby reducing information loss

  - See the paper for detailed formulation. The unified objective is **quantifiable** and derived by **variational principle**.

Empirical Analysis

- **Verification of Reduced Information Loss**



1. Utilizing proxy distributions **empirically reduces the information loss**

$$\mathcal{L}_{\text{NAT}} = \mathbb{E}_{Q(Z|X)} \mathcal{D}_{\text{KL}} \left[Q(T|Z,X) || P_\theta(T|Z,X) \right]$$

Empirical Analysis

- **Our Objective Strongly Correlated with BLEU**

Models	$\mathcal{L}_{\text{NAT}}$	$\hat{\mathcal{L}}_{\text{target}}$	$\hat{\mathcal{L}}_{\text{MPLE}}$	BLEU
Raw Data	4.41	-6.42	-2.01	11.79
KD	2.42	**-7.08**	-4.66	20.87
+ AXE (τ=1)	**0.78**	-5.13	-4.35	18.56
+ AXE (τ=5)	1.09	-6.34	-5.25	22.22
+ AXE (τ=10)	1.25	-6.50	-5.26	22.35
+ OaXE (10k)	1.03	-4.41	-3.38	15.00
+ OaXE (50k)	0.79	-5.84	-5.06	21.37
+ OaXE (300k)	0.83	-6.28	**-5.44**	**22.76**

Pearson's $|r| = 0.99$ for WMT14 En-De
$|r| = 0.96$ on WMT17 Zh-En

2. Our objective **jointly considers information loss and data distortion**, which **correlates well with generation performance**

Other Results & Analysis

- Our perspective can apply to many previous work in NAT learning, including
  - iterative NATs
  - latent variable models
  - CTC
  - DA-Transformer (our other work at ICML, which improves non-iterative NATs by 3 BLEU)

Other Results & Analysis

- Our perspective can apply to many previous work in NAT learning, including
  - iterative NATs
  - latent variable models
  - CTC
  - DA-Transformer (our other work at ICML, which improves non-iterative NATs by 3 BLEU)

- Our objective can guide the design of new training methods.
- About masking strategies, decoding methods ……

Thanks for Your Attention

Acknowledgement: Yuxuan Song

If you are interested, welcome to see our other paper at ICML2022!

Directed Acyclic Transformer for Non-Autoregressive Machine Translation

CoAI group, Tsinghua University

ByteDance AI Lab

Institute for AI Industry Research, Tsinghua University

University of California, Santa Barbara