IterativeVAE: Non-Autoregressive Neural Sequence Modeling by Iterative Refinement from Latent Space

刘泓尊，严宇康，唐李源，叶鲁斌

Introduction

在序列生成任务中，非自回归 Seq2Seq 模型 (NAR[1]) 无需迭代就可以高效生成目标句子，速度比通常使用的自回归模型 (AR) 快一个数量级。但是，因为多模态 (multimodality) 问题，一次并行且独立地生成所有词难以获得合理的结果，性能明显低于 AR 模型。对此，CMLM[2] 通过学习预测目标序列的空缺位置来建模单词之间的语义关系，以半自回归的方式迭代生成序列，达到了类似 AR 的效果；FlowSeq[3] 通过 Generative Flow 在隐空间建模复杂的目标序列分布，获得了较强的序列表征能力；

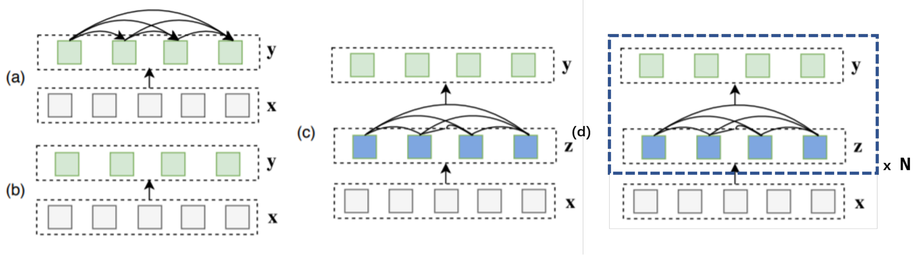

图 1: **(a)** AutoRegressive Models. **(b)** Non-AutoRegressive Models. **(c)** VAE/Flow-based Models. **(d)** Our Proposed Sequence Generation Models. x is the source, y is the target, and z are latent variables.

基于上面两种较为成功的 NAR 模型，我们提出了一种新的基于 Iterative Refinement 的 NAR: IterativeVAE. (1)IterativeVAE 通过在连续空间建模复杂隐变量，并且 (2) 逐层迭代修正，以期待生成更流畅、合理的结果; (3) 这种方法几乎适用于任何的 Seq2Seq 任务：我们在机器翻译 (De→En) 和开放域对话生成上评估了我们的模型，发现它可以在保证句子质量的同时大幅提高解码效率。(4)IterativeVAE 在生成序列质量上超过了 NAT Baseline，并且有比 CMLM 和 FlowSeq 更快的解码速度。

Method and Model

NAR 使用 target token 相互独立假设，在给定输入的情况下预测长度且并行生成所有 token:

$$p_{\mathcal{NA}}(Y|X;\theta) = p_L(T|x_{1:T};\theta)\cdot\prod_{t=1}^{T} p(y_t|x_{1:T};\theta) \tag{1}$$

而 VAE[3] 使用隐变量建模单词之间的条件依赖:

$$P_{\mathcal{VAE}}(y|x;\theta) = \int_z P_\theta(y|z,x)p_\theta(z|x)dz \tag{2}$$

在引入 Iterative Refinement 之后，Decoder Stack 的每层都将上一层的输出 y_{i-1} 作为输入 x, 并在连续空间 (词表维度) 建模新的隐变量，生成该层的 y_i 。通过逐层迭代，模型可以获得更强的隐空间表示。

IterativeVAE 在训练阶段逐层优化**证据下界 (ELBO)**, 并使用逐层减少的 KL_{weight} 控制 Hint 信息比重:

$$\mathcal{L}(\theta;x,y) = \mathbb{E}_{z_i\sim q_\theta(z_{i-1}|x,y)}[\log p_\theta(z_i|z_{i-1},x)] - KL\left[q_\theta(z_{i-1}|x,y)||p(z_{i-1}|x)\right] \tag{3}$$

其中 z_{i-1} 是第 i 层的输入，z_i 是第 i 层的输出，x 为源序列, y 为目标序列。

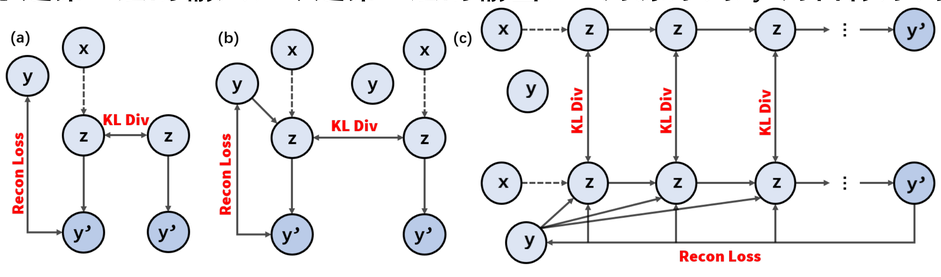

图 2: Graphical model representations. **(a)** Variational autoencoder (VAE). **(b)** Variational encoder-decoder (VED). **(c)** VED with iterative refinement (ours). **Dashed lines**: Encoding phase. **Solid lines**: Decoding phase.

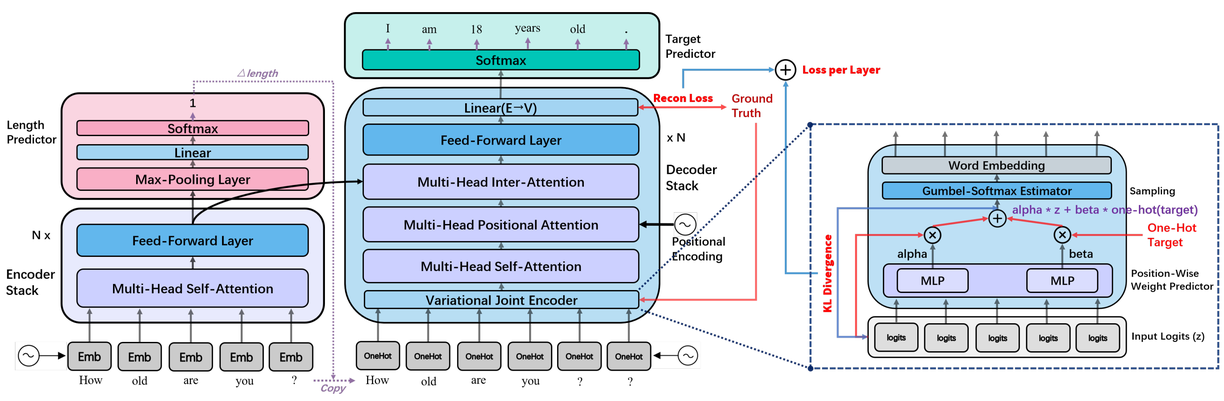

图 3: The architecture of the proposed model, where the black solid arrows represent differentiable connections and the purple dashed arrows are non-differentiable operations. Each sublayer inside the encoder and decoder stacks are standard, and uses both absolute and relative attention.

Experiments

在开放域对话任务上，我们使用 OpenSubtitles(1M pairs) 作为训练集，the Internet Movie Script Database(IMSDB) 作为验证和测试集 (just tokenized)；在机器翻译上，我们使用 IWSLT14(de→en)(use byte-pair encoding(BPE)) 训练和测试。

我们使用 BLEU1/BLEU2 评估对话生成质量, BLEU4 评估翻译质量, 使用 Distinct-1(dis-1), Distinct-2(dis-2) 评估 unigrams 和 bigrams 的多样性。我们还在数据集上 finetune GPT-2 来评估生成句子的 Language Model Score(LM Score)。Baselines 包括自回归 AR(use Standard Encode-Decoder Transformer), 半自回归 CMLM[2], VAE-Based FlowSeq[3] 和 Energy-based ENGINE.

参数配置为 $n_{layers}=6$, $n_{head}=8$, source/target embedding 共享, Base(default) models ($d_{model}/d_{hidden}=256/1024$), Large models ($d_{model}/d_{hidden}=512/2048$).

	#iter	BLEU1/2↑	dis-1/dis-2↑	LM Score↓	Avg.length	Latency(ms)↓/Speedup↑
AR-base(b=1)	N	4.89/3.44	0.814/0.806	4.54	9.8	420/2.15x
AR-base(b=4)	N	5.17/3.56	0.812/0.805	**4.33**	12.2	624/1.45x
AR-large(b=1)	N	4.96/3.48	**0.821/0.811**	4.61	10.3	610/1.48x
AR-large(b=4)	N	**4.96/3.51**	0.800/0.802	4.34	10.3	906/1.0x
NAT-base	1	1.51/0.33	0.011/0.017	5.87	8.8	84/10.78x
CMLM-base	1	2.1/0.53	0.277/0.639	5.70	10.2	81/11.18x
CMLM-base	4	2.09/0.71	0.334/0.662	4.93	10.2	291/1.44x
CMLM-large	1	2.65/0.66	0.156/0.482	5.65	12.5	191/4.74x
CMLM-large	4	**2.93/0.69**	**0.211/0.539**	5.65	12.5	774/1.17x
ENGINE-large	1	2.07/0.69	0.332/0.663	5.85	12.9	215/4.76x
FlowSeq-base	1	1.85/0.59	0.042/0.034	**4.50**	4.5	133/6.81x
IVAE-base(ours)	1(6)	1.86/0.61	0.143/0.313	4.81	7.9	**105/8.62x**
IVAE-large(ours)	1(6)	**1.89/0.61**	**0.149/0.334**	**4.75**	8.1	167/5.42x

表 1: Automatic Metrics Evaluation for Different Models on **IMSDB dataset**.

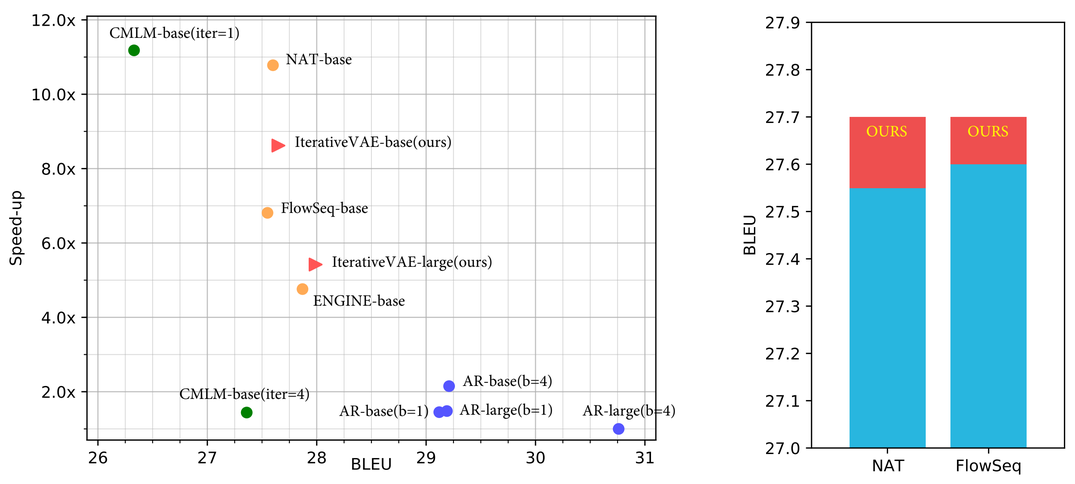

图 4: **Left**: The trade-off between speed-up and translation quality(BLEU) on dataset **IWSLT14(de→en)**. **Right**: Performance(BLEU) improved by our method compared with **NAR** and **FlowSeq** baseline on **IWSLT14**.

Layer Output	IMSDB	IWSLT14
Source	I told him there was a gas leak.	Nur so konnten wir beide zur Schule gehen.
Target	But he didn't listen to you.	That was the only way we could both get educated.
1	He he him with him him	That's the the way way can go go to.
2	He was him with his him	That's the the we we can go go to.
3	It was him at his work	That's the the we we could go go school.
4	He was him with his work	That's the the way i can go to school.
5	It was him at to work	That's the only way we could go to school.
6	It was him not to work	That's just the way we could go to school.

表 2: Iterative refinement sampling results on **IMSDB** and **IWSLT14** dataset.

Conclusion

我们提出了一种基于 Iterative Refinement 的 Variational Seq2Seq Model, 它可以在连续的隐空间建模复杂的特征分布，并使用迭代的方式不断改进生成结果。IterativeVAE 可以在保证 NAT 推理速度的同时提高生成质量，并且可以适用于机器翻译和对话生成等多个领域。未来的潜在发展包括深入研究模型的隐空间，从而获得更好的隐空间表示；以及引入 Masked Language Model 来进一步提高生成质量。

References

[1] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation, 2017.

[2] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models, 2019.

[3] Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. Flowseq: Non-autoregressive conditional sequence generation with generative flow. *EMNLP 2019*, 2019.