学界 | 非自回归神经机器翻译

2017-11-16 机器海岸线

作者：Jiatao Guy, James Bradbury, Caiming Xiong, Victor O.K. Liy& Richard Socher 等

机器海岸线编译

参与：方建勇

NON-AUTOREGRESSIVE
NEURAL MACHINE TRANSLATION

Jiatao Gu[†]*, James Bradbury[‡], Caiming Xiong[‡], Victor O.K. Li[†]& Richard Socher[‡]
[‡]Salesforce Research
{james.bradbury,cxiong,rsocher}@salesforce.com
[†]The University of Hong Kong
{jiataogu, vli}@eee.hku.hk

论文链接：

https://einstein.ai/static/images/pages/research/non-autoregressive-neural-mt.pdf

摘要：现有的神经机器翻译方法对先前生成的输出中的每个输出字进行调整。 我们引入一个模型来避免这种自回归属性并且产生并行输出，从而在推断过程中允许一个数量级的更低的延迟。 通过知识提炼，使用输入标记繁殖能力令牌作为潜在变量，以及策略梯度微调，相对于用作教师的自回归 Transformer 网络，花费仅为 2.0 BLEU 点的成本。 我们展示了与联讯战略三个方面相关的大量累积改进，并验证了我们在 IWSLT 2016 英语 - 德语和两个 WMT 语言对上的方法。 通过在推断时间并行采样繁殖能力，我们的非自回归模型在 WMT 2016 英语 - 罗马尼亚语上达到了 29.8 BLEU 的接近最先进的性能。

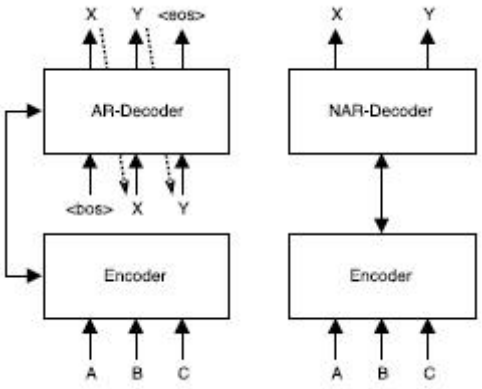

图 1：使用自回归和非自回归神经 MT 结构将 A B C 翻译为 X Y。后者可并行生成所有输出令牌。

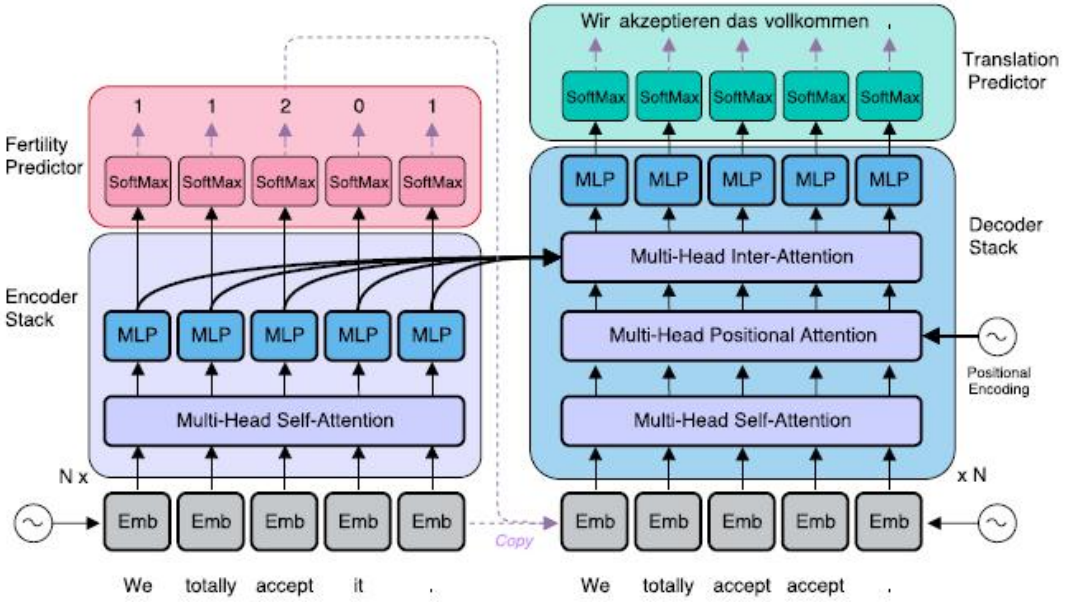

图 2：当 NAT 的体系结构，黑色的实线箭头表示可区分的连接，紫色的虚线箭头是不可区分的操作。 编码器和解码器堆栈内的每个子层还包括层规范化和剩余连接。

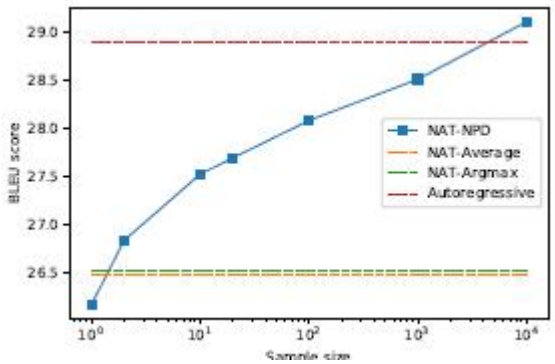

图 3：对于噪声并行解码，将 IWSLT 发展中的 BLEU 得分设置为样本大小的函数。 NPD 与两个样本之后的其他两个解码策略的性能相匹配，并且超过自回归教师的性能约 1000。

Models	WMT14		WMT16		IWSLT16		
	En→De	De→En	En→Ro	Ro→En	En→De	Latency	Speedup
NAT	17.35	20.62	26.22	27.83	25.20	39 ms	15.6×
NAT (+FT)	17.69	21.47	27.29	29.06	26.52	39 ms	15.6×
NAT (+FT + NPD $s = 10$)	18.66	22.41	29.02	30.76	27.44	79 ms	7.68×
NAT (+FT + NPD $s = 100$)	19.17	23.20	29.79	**31.44**	28.16	257 ms	2.36×
Autoregressive ($b = 1$)	22.71	26.39	31.35	31.03	28.89	408 ms	1.49×
Autoregressive ($b = 4$)	23.45	27.02	31.91	31.76	29.70	607 ms	1.00×

表 1：BLEU 在官方测试集上得分（WMT En-De 的 newstest2014 和 WMT En-Ro 的 newstest2016）或者 IWSLT 的开发集。 没有 NPD 的 NAT 模型使用 argmax 解码。 延迟时间被计算为在没有最小化的情况下解码单个句子的时间，在整个测试集上平均; 在单个 NVIDIA Tesla P100 上的 PyTorch 中执行解码。

Distillation		Decoder Inputs		+PosAtt	Fine-tuning			BLEU	BLEU (T)
b=1	b=4	+uniform	+fertility		+$\mathcal{L}_{\text{KD}}$	+$\mathcal{L}_{\text{BP}}$	+$\mathcal{L}_{\text{RL}}$		
				✓				≈ 2	
		✓		✓				16.51	
			✓	✓				18.87	
✓		✓		✓				20.72	
	✓	✓		✓				21.12	
✓			✓					24.02	43.91
✓			✓	✓				25.20	45.41
✓		✓		✓	✓	✓		22.44	
✓			✓	✓			✓	×	×
✓			✓	✓		✓		×	×
✓			✓	✓	✓	✓		25.76	46.11
✓			✓	✓	✓	✓	✓	**26.52**	**47.38**

表 2：在 IWSLT 开发集上的消融性能。 BLEU（T）是指由教师模型翻译的开发集版本上的 BLEU 分数。 An 表示微调导致该模型变得更糟。 当使用统一复制作为解码器输入时，提供接地真实目标长度。 所有型号都使用 argmax 解码。

Source:	politicians try to pick words and use words to shape reality and control reality , but in fact , reality changes words far more than words can ever change reality .
Target:	Politiker versuchen Worte zu benutzen , um die Realität zu formen und die Realität zu kontrollieren , aber tatsächlich verändert die Realität Worte viel mehr , als Worte die Realität jemals verändern könnten .
AR:	Politiker versuchen Wörter zu wählen und Wörter zur Realität zu gestalten und Realität zu steuern , aber in Wirklichkeit verändert sich die Realität viel mehr als Worte , die die Realität verändern können .
NAT:	Politiker versuchen , Wörter wählen und zu verwenden , um Realität zu formen und Realität zu formen , aber tatsächlich ändert Realität Realität viel mehr als Worte die Realität Realität verändern .
NAT+NPD:	Politiker versuchen , Wörter wählen und zu verwenden , um Realität Realität formen und die Realität zu formen , aber tatsächlich ändert die Realität Worte viel mehr als Worte jemals die Realität verändern können .
Source:	I see wheelchairs bought and sold like used cars .
Target:	Ich erlebe , dass Rollstühle gekauft und verkauft werden wie Gebrauchtwagen
AR:	Ich sehe Rollstühlen , die wie Autos verkauft und verkauft werden .
NAT:	Ich sehe , dass Stühle Stühle und verkauft wie Autos verkauft .
NAT+NPD:	Ich sehe Rollühle kauften und verkaufte wie Autos .

图 4：两个例子比较自回归（AR）和非自回归变换器产生的译文以及噪声并行译码的结果与样本大小 100. 重复的单词以灰色突出显示。

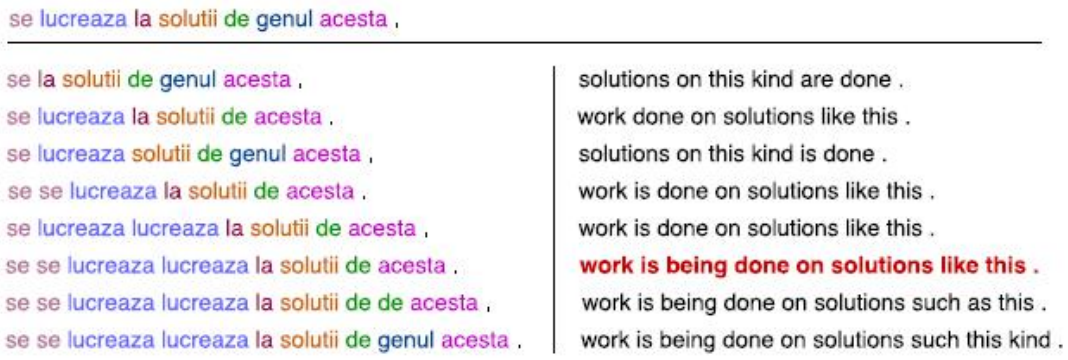

图 5：罗马尼亚语 - 英语示例翻译与噪声并行解码。 左边是来自编码器的八个采样生育力序列，用它们对应的解码器输入序列表示。 潜在变量的每个值都会导致不同的可能输出转换，如右图所示。 然后，自动回归变压器选择红色显示的最佳翻译，这个过程比直接使用它来产生输出要快得多。

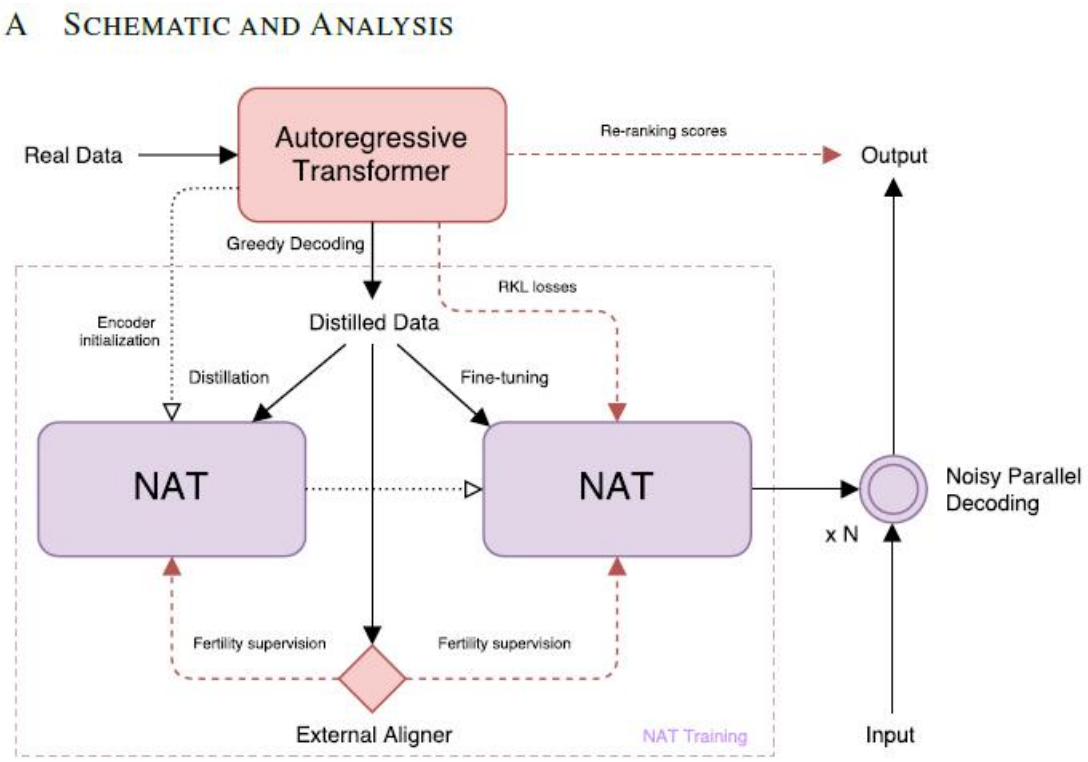

图 6：NAT 的训练和推断的示意结构。“提炼数据”包含由自回归模型解释的目标句子和地面真值源语句。

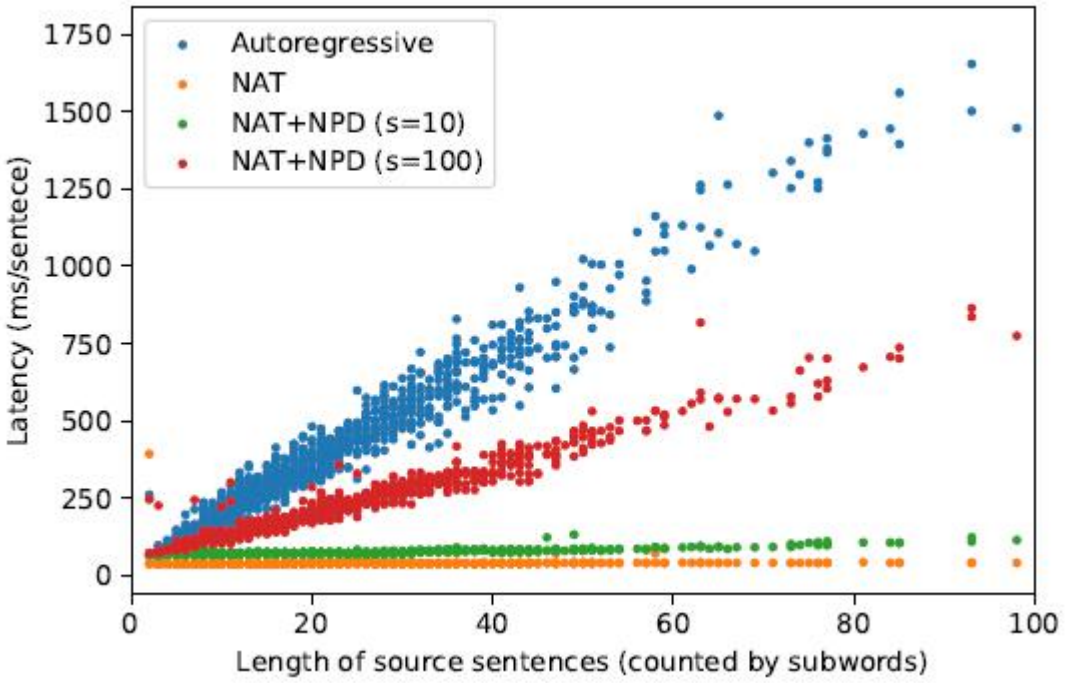

图 7：根据 IWSLT 开发中的每个句子，翻译等待时间（作为解释单个句子而不使用小批量处理的时间）计算，作为其长度的函数。 自回归模型在解码长度上具有线性延迟，而对于典型长度，NAT 的延迟几乎是恒

定的，即使对于样本大小为 10 的 NPD 也是如此。当使用样本大小为 100 的 NPD 时，并行性水平足够多 比饱和 GPU，再次导致线性延迟。

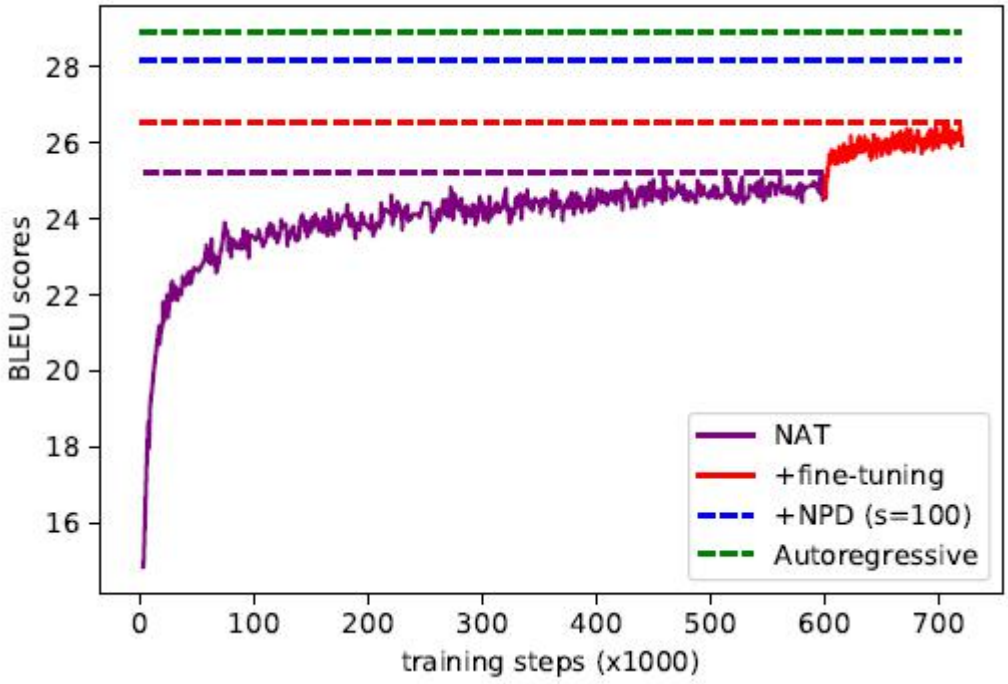

图 8：在 IWSLT 上学习用于训练和微调 NAT 的曲线。 BLEU 成绩正在发展中。

✂----------