中国科学院大学

University of Chinese Academy of Sciences

自然语言处理 机器翻译 阅读报告

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

2022 年 1 月 1 日

导师：胡玥 教授

姓名：	陈若愚
学号：	202118018629015
学院：	网络空间安全学院
专业：	计算机应用技术

Empower Distantly Supervised Relation Extraction with Collaborative Adversarial Training

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad,
Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
mikelewis@fb.com, yinhan@ai2incubator.com, naman@fb.com

Accepted by *ACL 2020*

ABSTRACT: 介绍了 BART，一种用于预训练序列到序列模型的去噪自编码器。通过（1）使用任意噪声函数来对文本进行加噪，并（2）学习模型以重建原始文本来训练 BART。它使用基于标准 Tranformer 的神经机器翻译架构，尽管它很简单，但可以看作是 BERT（由于双向编码器），GPT（具有从左至右解码器）以及许多其他最近的预训练方案的扩展。我们评估了多种加噪方法，发现通过随机改变原始句子的排列顺序并使用新的填充方案（其中文本段被单个 mask 标记替换）能获得最佳性能。当针对文本生成进行微调时，BART 特别有效，并且对于理解任务也很有效。在使用可比较的 GLUE 和 SQuAD 资源训练时，它与 RoBERTa 的性能相匹配，并在一系列对话生成，问答和摘要任务方面取得了最好的成果，且获得了多达 6 ROUGE 的收益。与机器翻译的反向翻译系统相比，仅对目标语言进行了预训练，BART 还提供了 1.1 个 BLEU 的提升。我们还报告了在 BART 框架内使用其他预训练方案的消融实验，以更好地衡量哪些因素最能影响下游任务性能。

1. Motivation

1.1 Problem

预测屏蔽字符的顺序以及替换屏蔽字符的可用上下文可以获得较大的收益。但是，这些方法通常专注于特定类型的下游任务（例如，跨度预测，生成等），从而限制了它们的适用性。

1.2 Contribution

这篇文章的主要贡献如下：

(1) 提出使用多种噪声破坏原文本，再将残缺文本通过序列到序列的任务重新复原的预训练任务

(2) BART 模型的提出解决了预训练模型编码器、解码器表征能力不一致的问题

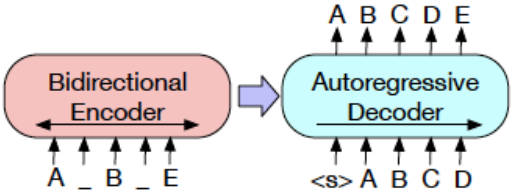
(3) 在 2019 年生成式 NLP 任务的榜单上刷新了多个 SOTA，验证了模型的有效性与先进性

1.3 Innovation

预训练有两个阶段（1）使用任意的噪声函数对文本进行加噪，以及（2）学习序列到序列模型以重建原始文本。通过强制模型对更复杂句子结构进行推理，并对输入进行更长的范围转换，从而扩展了 BERT 中的原始单词屏蔽和下一句预测目标。BART 模型堆叠在其他几个 transformer 层之上。这些层经过训练将外语本质上翻译为具有噪声的英语，然后通过 BART 进行传播，从而将 BART 作为预训练的目标端语言模型。

(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

图 1：BART 与 BERT 和 GPT 的示意图比较。

2. Model

BART 是一种去噪自编码器，可将具有噪声的文档映射到原始文档。它被实现为序列到序列模型，该模型包含在具有噪声的文本上进行双向编码的编码器，以及从左至右的自回归解码器。对于预训练，我们主要优化原始文档的负对数似然。

2.1 Architecture

BART 使用标准 Seq2Seq 的 transformer 架构，并与 GPT 类似，将 ReLU 激活函数修改为 GeLU，并以$\mathcal{N}(0,0.02)$初始化参数。对于我们的 Base 模型，我们使用

6 层编码器和解码器，对于我们的 Big 模型，我们则使用 12 层。BART 架构与 BERT 中使用的架构密切相关，但有以下**区别：（1）解码器的每一层还对编码器的最终隐藏层执行自注意力（如在 transformer 序列到序列模型中一样）；（2）BERT 在进行单词预测之前使用了一个附加的前馈网络，而 BART 则没有。**总体而言，BART 包含的参数比同等大小的 BERT 模型大约多了 10%。

2.2 Pre-training BART

对 BART 的训练是通过对文档加噪，然后优化重构损失（解码器输出与原始文档之间的交叉熵）来进行的。与针对特定加噪方案量身定制的现有去噪自编码器不同，BART 允许我们应用任何类型的文档加噪方式。在极端情况下，所有有关源的信息都丢失了，BART 相当于一种语言模型。

我们尝试了几种先前提出的新的加噪方法，但我们相信其他新替代方法具有更大的潜力。下面总结了我们使用的加噪方法，图 2 显示了示例。

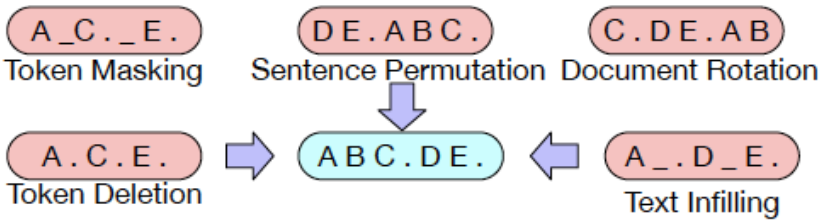


图 2：用来干扰我们实验的输入的变换。这些转换可以组合在一起。

Token Masking。与 BERT 相似，对随机字符进行采样并替换为[MASK]元素。

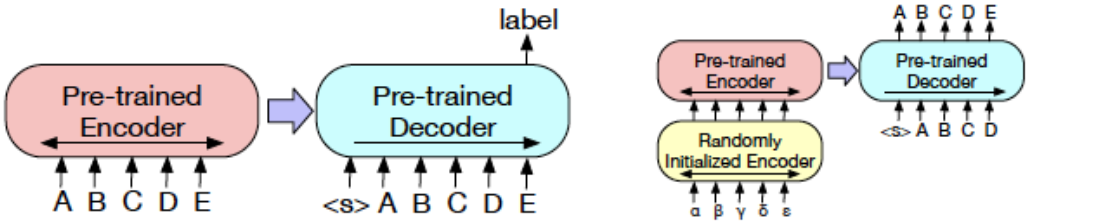
Token Deletion。随机字符将从输入中删除。与字符屏蔽相反，该模型必须确定哪些位置缺少输入。

Text Infilling。采样了多个文本跨度，跨度的长度来自泊松分布（$\lambda = 3$）。每个跨度都替换为一个[MASK]字符。0 长度跨度对应于[MASK]字符的插入。文本填充的灵感来自于 SpanBERT，但是 SpanBERT 采样了来自不同（固定几何）分布的跨度长度，并用完全相同长度的[MASK]字符序列替换了每个跨度。文本填充可以指导模型预测某个跨度中缺少多少字符。

Sentence Permutation。根据句号将文档分为多个句子，然后将这些句子随机排列。

Document Rotation。均匀地随机选择一个字符，然后旋转文档，使其从该字符开始。此任务训练模型以识别文档的开始。

3. Fine-tuning BART



(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

图 3：微调 BART 用于分类和翻译。

3.1 Sequence Classification Tasks

对于序列分类任务，将相同的输入馈入编码器和解码器，并将最后一个解码器字符的最终隐藏状态馈入新的线性分类器。此方法与 BERT 中的 CLS 字符相似；但是，我们时在末尾添加了额外的字符，以使解码器中字符的表示形式可以参与来自完整输入的解码器状态（图 3a）。

3.2 Token Classification Tasks

对于字符分类任务，例如 SQuAD 的答案终点分类，我们将完整文档输入编码器和解码器，并使用解码器的顶部隐藏状态作为每个单词的表示。该表示用于对字符进行分类。

3.3 Sequence Generation Tasks

由于 BART 具有自回归解码器，因此可以针对诸如生成式问答和摘要之类的序列生成任务直接进行微调。在这两个任务中，信息都是从输入中复制但受到操纵的，这与去噪预训练目标密切相关。此处，编码器输入是输入序列，解码器自动生成输出。

3.4 Machine Translation

我们还将探索使用 BART 来改进机器翻译解码器以翻译成英语。以前的工作显示可以通过合并预训练的编码器来改进模型，但是在解码器中使用预训练的语言模型的收益受到限制。我们证明，通过添加从 bitext 学习到的一组新的编码器参数，可以将整个 BART 模型（编码器和解码器）用作单个预训练解码器进行机器翻译（参见图 3b）。

更准确地说，我们将 BART 的编码器嵌入层替换为新的随机初始化的编码器。该模型是端到端训练的，它训练了新的编码器以将其他语言的词映射到输入中，BART 可以将其去噪为英语。新的编码器可以使用与原始 BART 模型不同的词表。

我们分两步训练源编码器，在两种情况下都从 BART 模型的输出反向传播交叉熵损失。第一步，我们冻结大多数 BART 参数，仅更新随机初始化的源编码器，BART 位置嵌入以及 BART 编码器第一层的自注意输入投影矩阵。第二步，我们训练所有模型参数进行少量迭代。

4. Comparing Pre-training Objectives

BART 在预训练期间支持比以前的工作更广泛的噪声方案。 我们使用基本大小模型（6 个编码器和 6 个解码器层，隐藏大小为 768）比较了一系列选项，在我们将在 x5 中进行全面大规模实验的任务的代表性子集上进行评估。

4.1 Comparison Objectives

Language Model： 与 GPT 类似，我们训练一个从左到右的 Transformer 语言模型。该模型等效于 BART 解码器，不需要交叉注意。

Permuted Language Model： 基于 XLNet，我们对 1/6 的 token 进行采样，并自回归随机顺序生成它们。为了与其他模型保持一致，我们没有实现相对位置嵌入或跨 XLNet 片段的注意。

Masked Language Model： 在 BERT 之后，我们将 15%的 token 替换为[MASK]符号，并训练模型独立预测原始 token。

Multitask Masked Language Model： 在 UniLM 中，我们使用额外的自我注意面具训练掩码语言模型。自我注意掩码随机选择如下比例:1/6 从左到右，1/6 从右到左，1/3 取消掩码，1/3 前 50%的标记取消掩码，其余的从左到右。

Masked Seq-to-Seq： 受 MASS 的启发，我们将包含 50%的令牌的跨度进行掩码，并将一个序列训练为序列模型来预测掩码 token。

4.2 Tasks

- **SQuAD**
- **MNLI**
- **ELI5**
- **XSum**
- **ConvAI2**

- **CNN/DM**

4.3 Results

结果如表 1 所示，证明了 BART 的优势，state-of-art。

表 1：预训练目标的比较，包括受 BERT、MASS、GPT、XLNet 和 UniLM 启发的方法。所有模型的大小都与 BERT Base 相似，并且在相同的数据上训练了 1M 步。底部两个块中的条目使用相同的代码库在相同的数据上进行训练，并使用相同的程序进行微调。第二个块中的条目受到先前工作中提出的预训练目标的启发，但已被简化以专注于评估目标（参见 4.1 节）。不同任务的性能差异很大，但带有文本填充的 BART 模型表现出最一致的强大性能。

Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

5. Large-scale Pre-training Experiments

5.1 Experimental Setup

预训练一个大模型，每个编码器和解码器都有 12 层，隐藏大小为 1024。按照 RoBERTa 的方法，我们使用 8000 个批量，并将模型训练为 500000 步。文档使用与 GPT-2 相同的字节对编码进行标记。根据第 4.3 节的结果，我们使用了文本填充和句子排列的组合。我们在每个文档中隐藏 30%的 token，并对所有句子进行置换。尽管句子排列仅在 CNN/DM 摘要数据集上显示显著的附加增益，但我们假设，更大的预训练模型可能更能从该任务中学习。为了帮助模型更好地拟合数据，我们在最后 10%的训练步骤中禁用了 dropout。

5.2 Discriminative Tasks

表 3 和表 2 比较了 BART 和几个最近的研究过的关于 SQuAD 和 GLUE 任务的方法的性能。

最直接可比较的基准是 RoBERTa，它使用相同的资源进行了预训练，但目标不同。总体而言，BART 的表现类似，在大多数任务上，模型之间只有很小的差异。这表明 BART 在生成任务上的改进不会以分类性能为代价。

表 2： GLUE 任务上的大型模型的结果。BART 的性能与 RoBERTa 和 XLNet 相当，这表明在有区别的任务上，BART 的单向解码器层不会降低性能。

	MNLI m/mm	SST Acc	QQP Acc	QNLI Acc	STS-B Acc	RTE Acc	MRPC Acc	CoLA Mcc
BERT	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

表 3： BART 在问题回答上给出了与 XLNet 和 RoBERTa 相似的结果。

	SQuAD 1.1 EM/F1	SQuAD 2.0 EM/F1
BERT	84.1/90.9	79.0/81.8
UniLM	-/-	80.5/83.4
XLNet	**89.0**/94.5	86.1/88.8
RoBERTa	88.9/**94.6**	**86.5/89.4**
BART	88.8/**94.6**	86.1/89.2

5.3 Generation Tasks

还试验了几个文本生成任务。作为从输入到输出文本的标准序列到序列模型，BART 被微调。在细调过程中，我们使用标签平滑交叉熵损失(Pereyra 等人，2017)，并将平滑参数设置为 0.1。在生成过程中，我们设置波束尺寸为 5，在波束搜索中去除重复的三元组，并在验证集上使用 min-len、max-len、length penalty 对模型进行调整

为了与最先进的摘要进行比较，我们展示了两个摘要数据集的结果，CNN/DailyMail 和 XSum，它们有不同的属性(表 4)。

我们还进行了人工评估(表 5)。研究者被要求为一篇文章从两个摘要中选择一个更好的。其中一个摘要来自 BART，另一个是人类参考或来自 BERTSUMEXTABS 模型的公开输出。与自动化的度量一样，BART 显著优于之前的工作。然而，它在这个任务上还没有达到人类的性能。

5.4 Translation

我们还评估了 WMT16 罗马尼亚语-英语的表现，并补充了 Sennrich 等人(2016)的反向翻译数据。我们使用 6 层 Transformer 源 Encoder 将罗马尼亚语映射成

BART 能够去噪的表示形式，并遵循节 3.4 中引入的方法。实验结果如表 8 所示，效果非常显著。

表 4： 两个标准摘要数据集的结果。 BART 在任务和所有指标（包括基于大规模预训练的指标）的汇总方面均优于之前的工作。

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
ROBERTASHARE (Rothe et al., 2019)	40.31	18.91	37.62	41.45	18.79	33.90
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

表 5： 人类对 XSum 的评价:BART 总结比前人的总结更受欢迎，但不比人类写的参考总结好。

	XSum	
	Prefer BART	Prefer Baseline
vs. BERTSUMEXTABS	73%	27%
vs. Reference	33%	67%

表 6： BART 在会话响应生成方面的表现优于以前的工作。基于 ConvAI2 的官方 token 生成器，对扰动进行重新正规化。

	ConvAI2	
	Valid F1	Valid PPL
Seq2Seq + Attention	16.02	35.07
Best System [2]	19.09	17.51
BART	**20.72**	**11.85**

表 7： BART 在具有挑战性的 ELI5 抽象问题回答数据集上取得了最先进的结果。

	ELI5		
	R1	R2	RL
Best Extractive	23.5	3.1	17.5
Language Model	27.8	4.7	23.1
Seq2Seq	28.3	5.1	22.8
Seq2Seq Multitask	28.9	5.4	23.1
BART	**30.6**	**6.2**	**24.3**

表 8： WMT 16 RO-EN 上基线和 BART 的 BLEU 分数增加了反向翻译数据。 BART 通过使用单语英语预训练改进了强大的反向翻译基线。

	RO-EN
Baseline	36.80
Fixed BART	β6.29
Tuned BART	**37.96**

6. Qualitative Analysis

表 9 显示了 BART 生成的具有代表性的示例总结，说明了它的主要优点和缺点。示例取自 WikiNews 在建立训练前语料库之后发表的文章，以消除描述的事件出现在模型训练数据中的可能性。在总结之前删除了文章的第一句话，所以不容易对文档进行提取总结。不出所料，模型输出是流利且语法正确的英语。然而，输出

也是高度抽象的，很少有复制的短语。摘要通常是事实准确的，并将来自整个输入文档的支持性证据与背景知识相结合(例如，正确填写姓名，或推断 PG&E 在加州运营)。在第一个例子中，推断鱼类正在保护珊瑚礁不受全球变暖的影响需要一个重要的推论。然而，发表在《科学》杂志上的这项研究并没有得到来源的支持，而且，一般来说，该模型的主要局限是对不支持的信息产生幻觉的倾向。这些样本表明，BART 的前训练学习了自然语言理解和生成的强大结合。

表 9：来自维基新闻文章上 XSum 调整的 BART 模型的示例摘要。为清楚起见，仅显示了来源的相关摘录。 摘要结合了整篇文章中的信息和先验知识。

Source Document (abbreviated)	BART Summary
The researchers examined three types of coral in reefs off the coast of Fiji ... The researchers found when fish were plentiful, they would eat algae and seaweed off the corals, which appeared to leave them more resistant to the bacterium Vibrio coralliilyticus, a bacterium associated with bleaching. The researchers suggested the algae, like warming temperatures, might render the corals' chemical defenses less effective, and the fish were protecting the coral by removing the algae.	Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.
According to Syrian state media, government forces began deploying into previously SDF controlled territory yesterday. ... On October 6, US President Donald Trump and Turkish President Recep Tayyip Erdoan spoke on the phone. Then both nations issued statements speaking of an imminent incursion into northeast Syria ... . On Wednesday, Turkey began a military offensive with airstrikes followed by a ground invasion.	Syrian government forces have entered territory held by the US-backed Syrian Democratic Forces (SDF) in response to Turkey's incursion into the region.
This is the first time anyone has been recorded to run a full marathon of 42.195 kilometers (approximately 26 miles) under this pursued landmark time. It was not, however, an officially sanctioned world record, as it was not an "open race" of the IAAF. His time was 1 hour 59 minutes 40.2 seconds. Kipchoge ran in Vienna, Austria. It was an event specifically designed to help Kipchoge break the two hour barrier.	Kenyan runner Eliud Kipchoge has run a marathon in less than two hours.
PG&E stated it scheduled the blackouts in response to forecasts for high winds amid dry conditions. The aim is to reduce the risk of wildfires. Nearly 800 thousand customers were scheduled to be affected by the shutoffs which were expected to last through at least midday tomorrow.	Power has been turned off to millions of customers in California as part of a power shutoff plan.

7. Summary

这篇文章属于工程性问题，我们引入了 BART，这是一种预训练方法，可以学习将损坏的文档映射到原始文档。BART 在判别任务上的表现与 RoBERTa 相当，

并在多个文本生成任务上取得了最新的最新成果。未来的工作应该探索破坏预训练文档的新方法，也许可以针对特定的最终任务定制它们。方法简单，效果出众。

通过阅读此篇文章可以看出对 BERT 使用上是存在许多技巧的，并且对于预训练的第三范式方面是研究的热点。此外，对于机器翻译任务可使用 BERT 可以事半功倍。