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

Bart (2).pdf

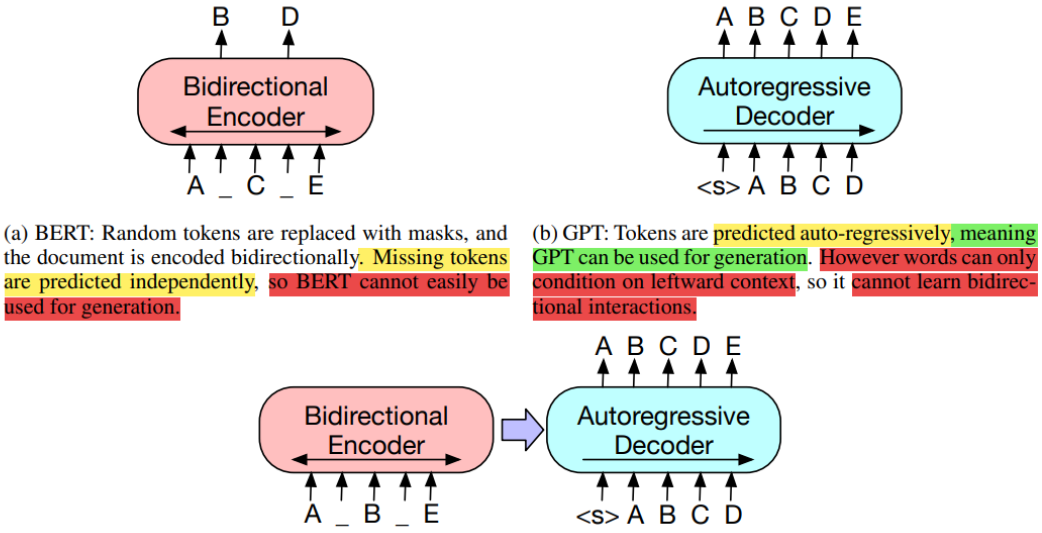
Abstract

- We present BART, a denoising autoencoder for pretraining sequence-to-sequence models.
- BART is trained by
    1. corrupting text with an arbitrary noising function
    2. learning a model to reconstruct the original text
- It uses a standard Tranformer-based neural machine translation architecture which

- despite its simplicity, can be seen as generalizing
  - **BERT** (due to the bidirectional encoder)
  - GPT (with the left-to-right decoder)
  - and many other more recent pretraining schemes
- BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks

1 Introduction

- Self-supervised methods have achieved remarkable success in a wide range of NLP tasks
- The most successful approaches have been variants of masked language models
- which are **de**noising autoencoders that are trained to reconstruct text where a random subset of the words has been masked out.
- Recent work has shown gains by improving
  - the distribution of masked tokens
  - the order in which masked tokens are predicted
  - the available context for replacing masked tokens
- However, these methods typically focus on particular types of end tasks (limiting their applicability)
  - span prediction
  - generation
- In this paper, we present BART
- **a model combining Bidirectional and Auto-Regressive Transformers.**
- BART is a denoising autoencoder built with a sequence-to-sequence model that is applicable to a very wide range of end tasks
- Pretraining has two stages
  1. text is corrupted with an arbitrary **noising function**
  2. a sequence-to-sequence model is learned to **reconstruct** the original text



(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 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.

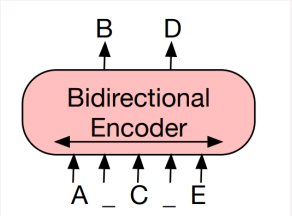
Figure 1: A schematic comparison of BART with BERT (Devlin et al., 2019) and GPT (Radford et al., 2018).

- A key advantage of this setup is the noising flexibility
- arbitrary transformations can be applied to the original text
- BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks.
- BART also opens up new ways of thinking about fine tuning.
- We present a new scheme for machine translation where a BART model is stacked above a few additional transformer layers.
- These layers are trained to essentially translate the foreign language to noised English
    - by propagation through BART
- thereby using BART as a pre-trained target-side language model.

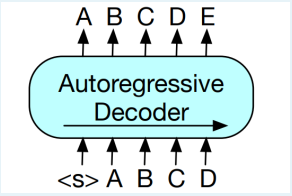
2. Model

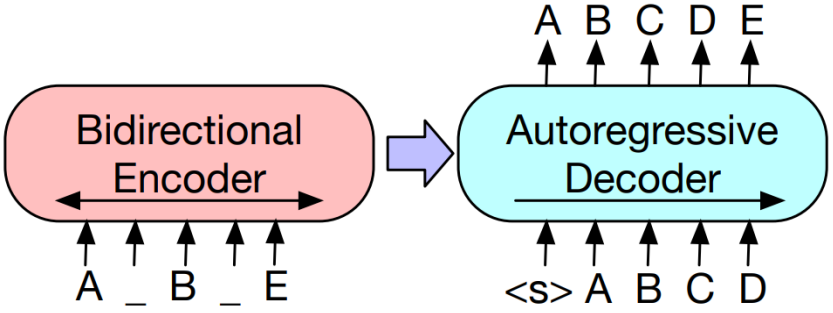
BART is a denoising autoencoder that maps a corrupted document to the original document it was derived from

- It is implemented as a
  - sequence-to-sequence model with a bidirectional encoder over corrupted text



  - a left-to-right autoregressive decoder.





- For pre-training, we optimize the negative log likelihood of the original document.

2.1 Architecture

- BART uses the standard sequence-to-sequence Transformer architecture from ***attention is all you need***
  - except, following GPT, that we modify $ReLU$ activation functions to $GeLUs$
  - initialise parameters from $N(0, 0.02)$. For our base model
  - using $6\ layers$ in the encoder and decoder
  - for the large model we use $12\ layers$ in each.

- The architecture is closely related to that used in BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- with the following differences
    1. each **layer** of the **decoder** additionally **performs cross-attention** over the final hidden layer of the encoder (as in the transformer sequence-to-sequence model);
    2. BERT uses an additional feed-forward network before word prediction , which BART does not
- BART contains roughly 10% **more parameters** than the equivalently sized BERT model.

2.2 Pre-training BART

> 🌟 BART is trained by corrupting documents and then optimizing a reconstruction loss—the cross-entropy between the decoder's output and the original document

- Unlike **existing denoising autoencoders**, which are **tailored to specific noising schemes**
- BART allows us to apply any type of document corruption
- In the extreme case, where all information about the source is lost, BART is equivalent to a language model.
- **Token Masking** Following BERT
- **Token Deletion** Random tokens are deleted from the input. In contrast to token masking, the model must decide which positions are missing inputs
- **Text Infilling**
    - A number of text spans are sampled
    - with span lengths drawn from a **Poisson distribution** $(\lambda = 3)$.
    - Each span is replaced with a single $[MASK]$ token.
        - 0-length spans correspond to the insertion of $[MASK]$tokens.
        - Text infilling is inspired by SpanBert

    - **Text infilling teaches the model to predict how many tokens are missing from a span.**

- **Sentence Permutation**
    - A document is **divided into sentences based on full stops**, and these sentences are shuffled in a random order.
- **Document Rotation**
    - A token is chosen uniformly at random
    - The document is rotated so that it begins with that token.
    - This task trains the model to **identify the start of the document.**

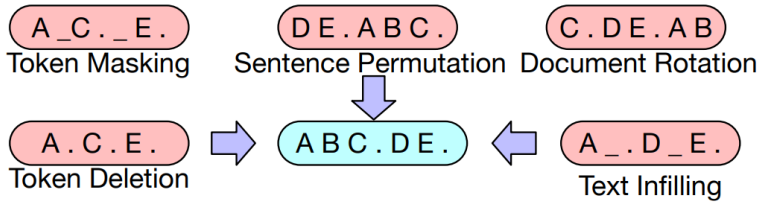


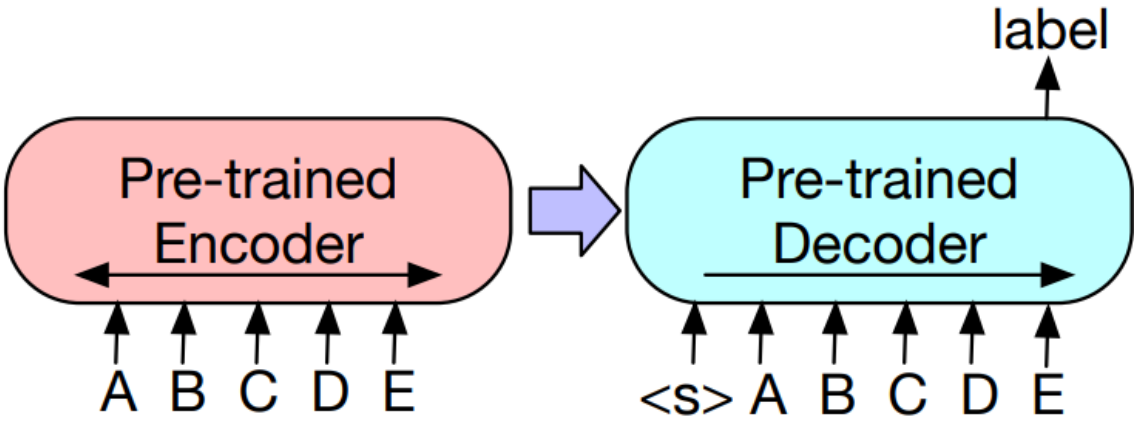
Figure 2: Transformations for noising the input that we experiment with. These transformations can be composed.

3. Fine-tuning BART

- The representations produced by BART can be used in several ways for downstream applications.

3.1 Sequence Classification Tasks

- For sequence classification tasks
    - the same input is fed into the encoder and decoder
    - the final hidden state of the final decoder token is **fed into** new multi-class linear classifier
- This approach is related to the CLS token in BERT
- however we **add the additional token to the end** so that **representation for the token in the decoder can attend to decoder states from the complete input**



(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.

3.2 Token Classification Tasks

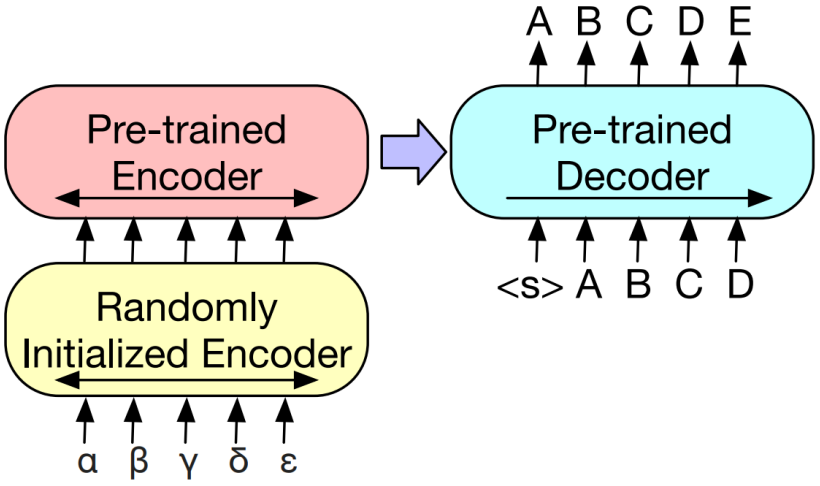
- For token classification tasks, such as answer endpoint classification for SQuAD
- we feed the complete document into the encoder and decoder.
- And use the top hidden state of the decoder as a representation for each word.
- This representation is used to classify the token

3.3 Sequence Generation Tasks

- Because BART has an autoregressive decoder
  - so it can be directly fine tuned for sequence generation tasks such
  - as **abstractive question answering** and **summarization**
  - In both of these tasks
    - information is copied from the input but manipulated
    - which is closely related to the denoising pre-training objective.
  - Here, the encoder input is the input sequence, and the decoder generates outputs autoregressively

3.4 Machine Translation

- We also explore using BART to improve machine translation decoders for translating into English.
- We show that it is possible to use the entire BART model (both encoder and decoder)
    - by adding a new set of encoder parameters that are learned from bitext
- More precisely, we replace BART's encoder embedding layer with a new randomly initialized encoder.



(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.

- The model is trained end-to-end
- which trains the **new encoder to map foreign** words **into an input that BART can de-noise to English**
- The new encoder **can use a separate vocabulary** from the original BART model
- We train the source encoder in two steps, in both cases backpropagating the cross-entropy loss from the output of the BART model
    1. we freeze❄️🧊 most of BART parameters and only update
        a. randomly initialized source encoder
        b. the BART positional embeddings
        c. and the self-attention input projection matrix of BART's encoder first layer
    2. we train all model parameters for a small number of iterations.

4. Comparing Pre-training Objectives

- BART supports a much wider range of noising schemes during pre-training than previous work
- We compare a range of options using base-size models
  - ($6\ encoder$ and $6\ decoder$ layers, with a $hidden\ size\ of\ 768$)

4.1 Comparison Objectives

> **we compare our implementations with published numbers from BERT, which was also trained for 1M steps on a combination of books and Wikipedia data. We compare the following approaches:**

- **Language Model**
  - Similarly to GPT (2018)
  - we train a left-to-right Transformer language model.
  - This model is equivalent to the BART decoder, without cross-attention.
- Permuted Language Model
  - based on XLNet
  - they sampled 1/6 of the tokens
  - and generate them in a random order autoregressively
  - they don’t implement
    - the relative positional embeddings or attention across segments from XLNet.
- Masked Language Model Following BERT
- Multitask Masked Language Model
  - we train a Masked Language Model with additional self-attention masks.
  - Self attention masks are chosen randomly in with the follow proportions:
    - 1/6 left-to-right
    - 1/6 right-to-left
    - 1/3 unmasked

        - 1/3 with the first 50% of tokens unmasked
        - and a left-to-right mask for the remainder
- Masked Seq-to-Seq
    - mask a span containing 50% of tokens, and train a sequence to sequence model to predict the masked tokens.

4.2 Tasks

- SQuAD
- XSum
    - a news summarization dataset with highly abstractive summaries.
- CNN/DM
    - a news summarization dataset. Summaries here are typically closely related to source sentences
- MNLI
    - a bitext classification task to predict whether one sentence entails another
- ELI5
    - a long-form abstractive question answering dataset.
    - Models generate answers conditioned on the concatenation of a question and supporting documents.
- ConvAI2
    - a dialogue response generation task, conditioned on context and a persona.

5. Large-scale Pre-training Experiments

5.3 Generation Tasks

- **Summarization**
- we present results on two summarization datasets, CNN/DailyMail and XSum.
- Summaries in the CNN/DailyMail tend to resemble source sentences.
    - **Extractive models** do well here, and even the baseline of the first-three source sentences is highly competitive.

- BART outperformed all existing work.(in that time)

- In contrast, XSum is highly **abstractive , and extractive models perform poorly.**
- BART outperforms the best previous work, which leverages BERT, by roughly 6.0 points on all ROUGE metrics

	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
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

Table 3: Results on two standard summarization datasets. BART outperforms previous work on summarization on two tasks and all metrics, with gains of roughly 6 points on the more abstractive dataset.

- on the CNN/DM summarization dataset, we hypothesised that larger pre-trained models may be better able to learn from this task. To help the model better fit the data, we disabled dropout for the final 10% of training steps.

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.

Table 7: Example summaries from the XSum-tuned BART model on WikiNews articles. For clarity, only relevant excerpts of the source are shown. Summaries combine information from across the article and prior knowledge.

- Deletion appears to outperform masking on generation tasks.
- Performance of pre-training methods varies significantly across tasks
- Left-to-right pre-training improves generation
    - The Masked Language Model and the Permuted Language Model perform less well than others on generation
- Bidirectional encoders are crucial for SQuAD
- The pre-training objective is not the only important factor
- Pure language models perform best on ELI5

- BART achieves the most consistently strong performance. With the exception of ELI5, BART models using text-infilling perform well on all tasks