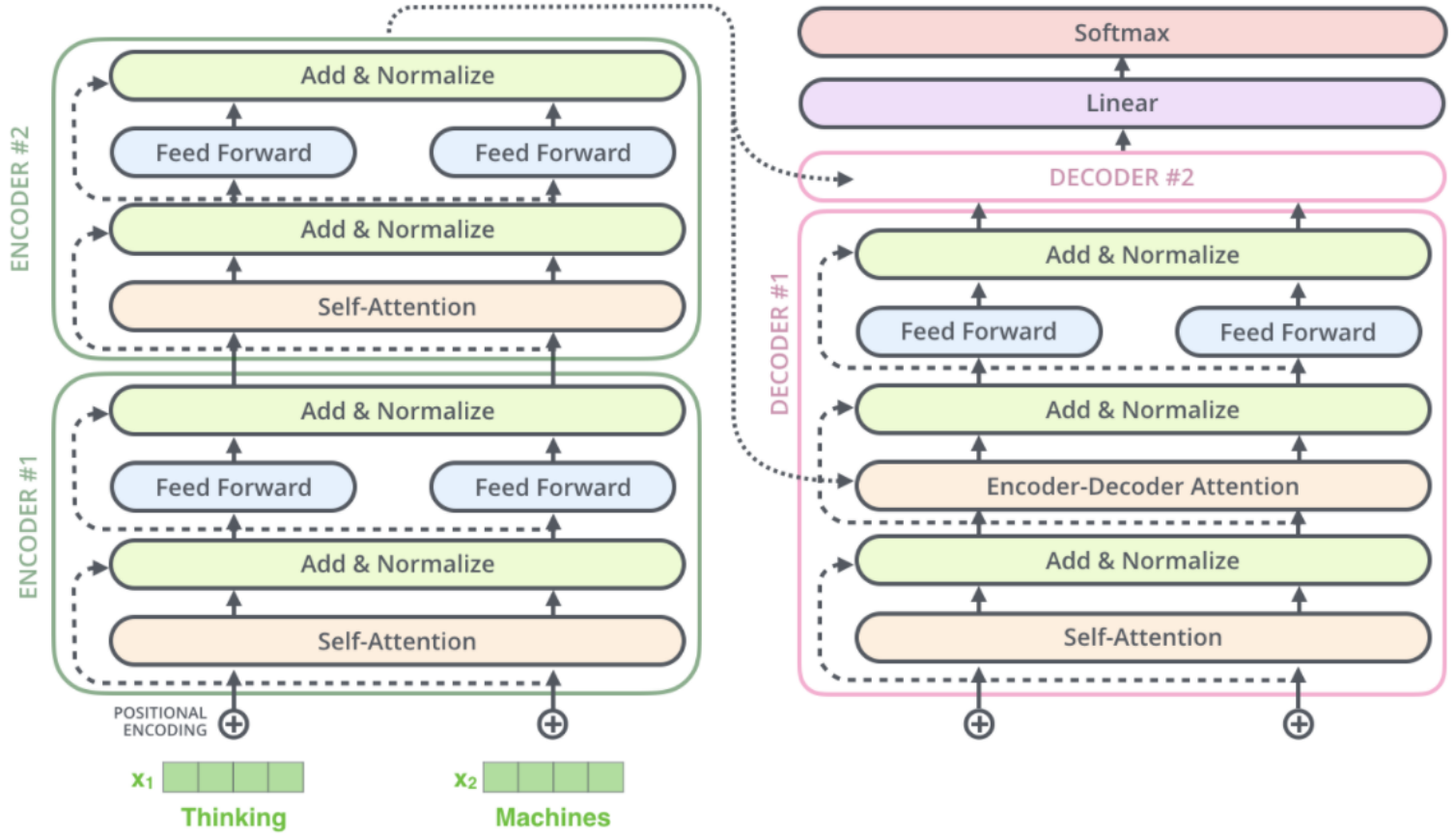
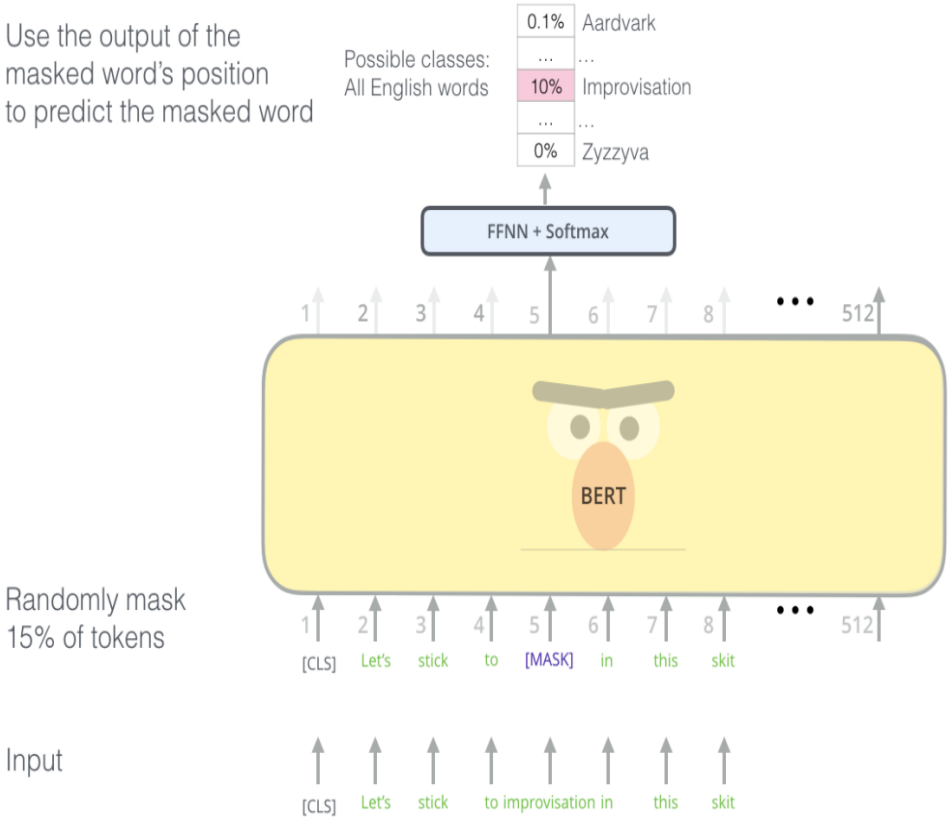
BART

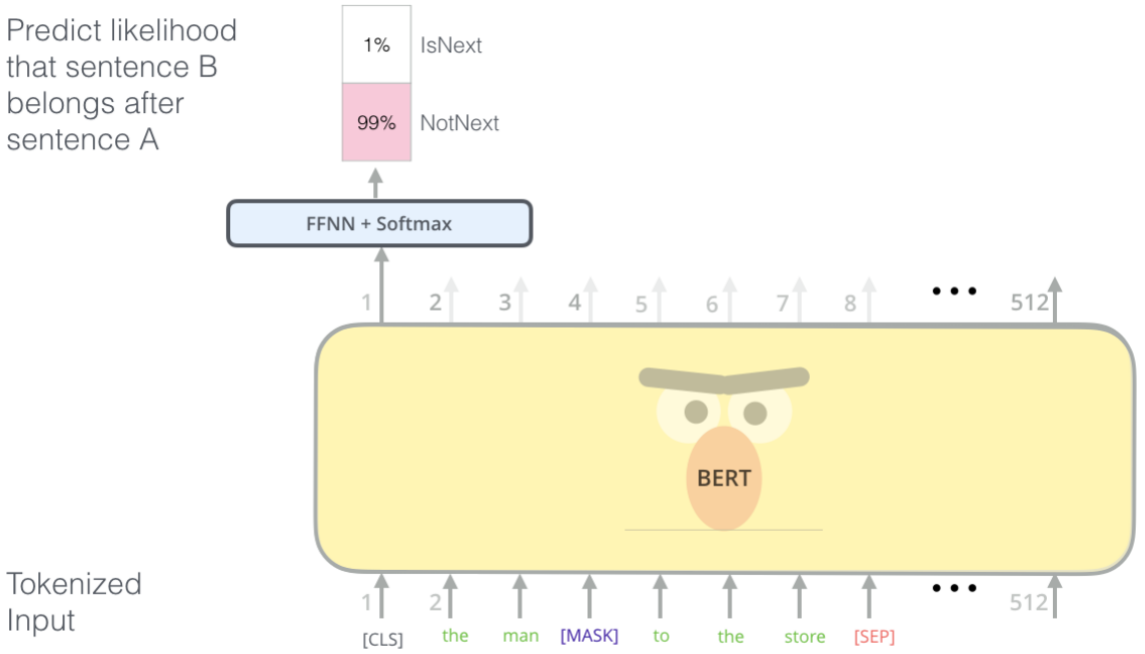
Mike Lewis et. al.

Transformer

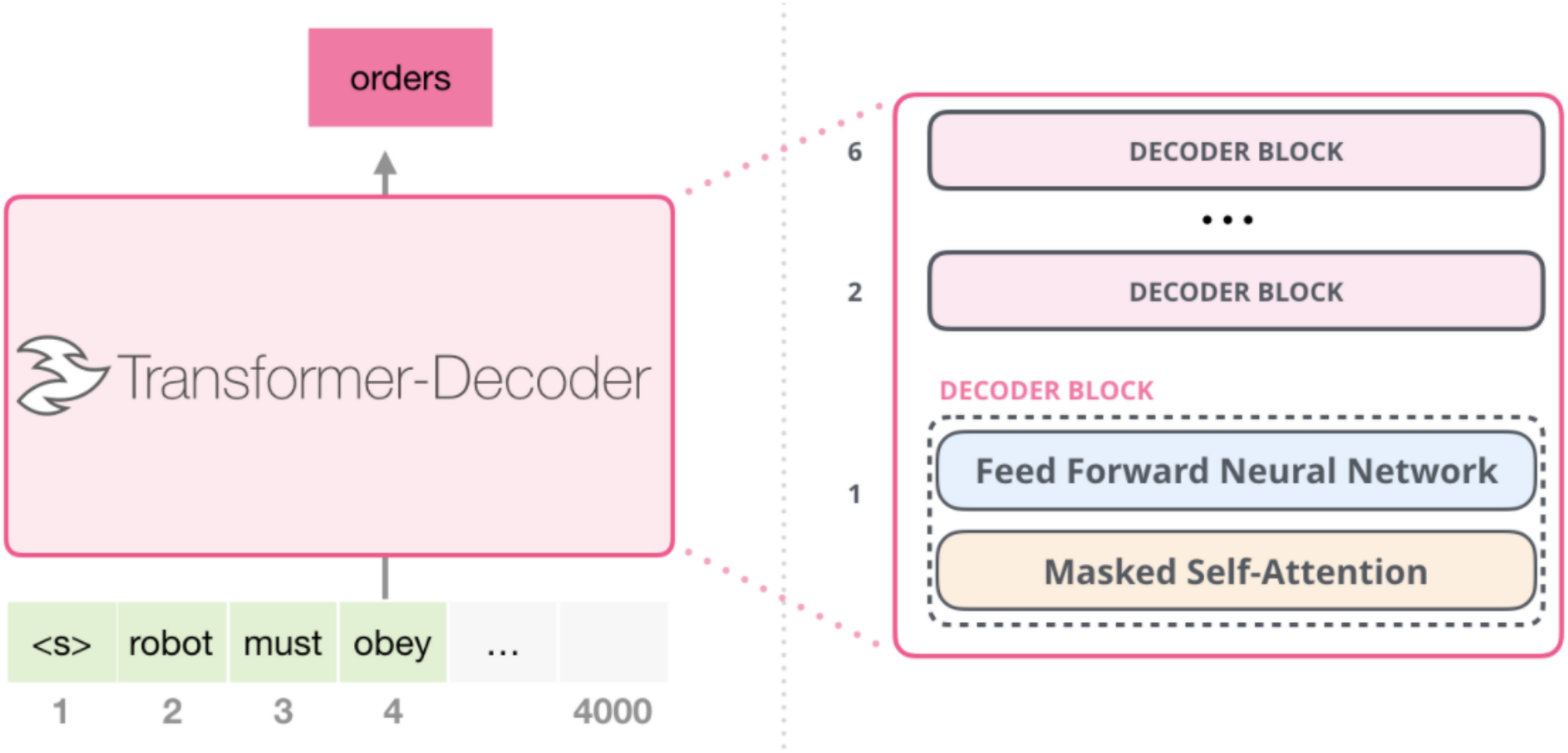


BERT


Use the output of the masked word's position to predict the masked word
Possible classes: All English words
0.1% Aardvark
...
10% Improvisation
...
0% Zyzzyva
FFNN + Softmax
1 2 3 4 5 6 7 8 ... 512
BERT
Randomly mask 15% of tokens
1 2 3 4 5 6 7 8 ... 512
[CLS] Let's stick to [MASK] in this skit
Input
[CLS] Let's stick to improvisation in this skit


Predict likelihood that sentence B belongs after sentence A
1% IsNext
99% NotNext
FFNN + Softmax
1 2 3 4 5 6 7 8 ... 512
BERT
Tokenized Input
1 2 ... 512
[CLS] the man [MASK] to the store [SEP]

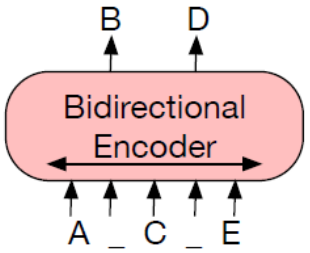
GPT


orders
Transformer-Decoder
<s>
robot
must
obey
…
1
2
3
4
4000
6
2
1
DECODER BLOCK
• • •
DECODER BLOCK
DECODER BLOCK
Feed Forward Neural Network
Masked Self-Attention

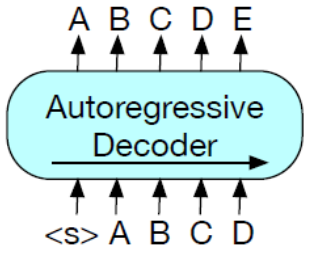
BART

- Denoising autoencoder for pretraining sequence-to-sequence models.
- Trained by corrupting text with an arbitrary noising function, and learning a model to reconstruct the original text.
- Uses a standard Tranformer-based neural machine translation architecture
  - Generalizing BERT , GPT

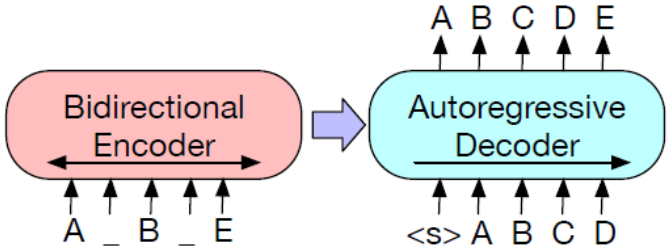
BART



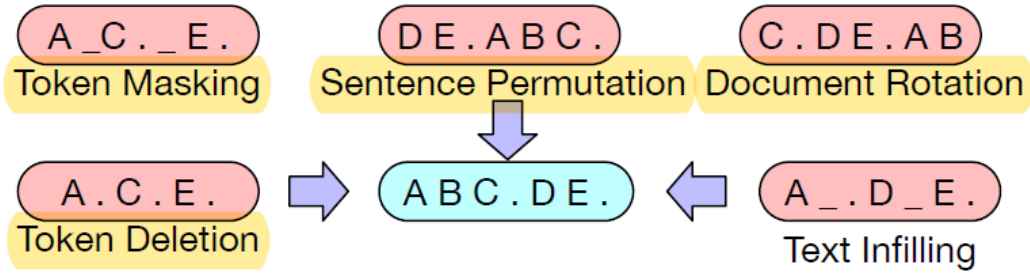
(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(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.

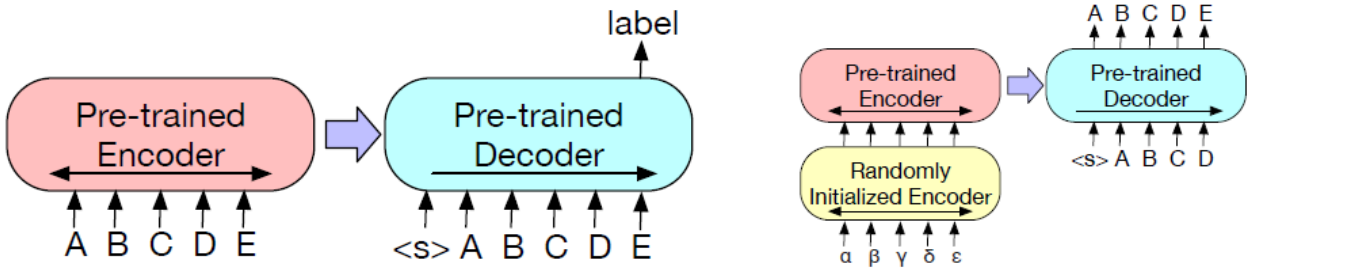


Pre-training BART

- Trained by corrupting documents and then optimizing a reconstruction loss—the cross-entropy between the decoder's output and the original document.
- BART allows us to apply any type of document corruption.

- Corruption schemes:
- Token Masking - random tokens are sampled and replaced with [MASK] elements.
- Token Deletion - Random tokens are deleted from the input.
- 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, and the document is rotated so that it begins with that token.
- Text Infilling - A number of text spans are sampled, with span lengths drawn from a Poisson distribution. Each span is replaced with a single [MASK] token. 0-length spans correspond to the insertion of [MASK] tokens.

Fine tuning in 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.

Figure 3: Fine tuning BART for classification and translation.

Fine tuning BART

- Sequence Classification Tasks - Same input is fed into the encoder and decoder, and the final hidden state of the final decoder token is fed into new multi-class linear classifier. Additional token to the end.
- Token Classification Tasks - feed the complete document into the encoder and decoder, use the top hidden state of the decoder as a representation for each word.
- Sequence Generation Tasks - encoder input is the input sequence, and the decoder generates outputs autoregressively.
- Machine Translation –
  - BART’s encoder embedding layer with a new randomly initialized encoder
  - Two stages : Freeze most of BART parameters and only update the randomly initialized source encoder, train all model parameters for a small number of iterations.

Results

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**

Results

	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**

Table 4: Results on two standard summarization datasets. BART outperforms previous work on summarization on both tasks and all metrics, including those based on large-scale pre-training.

	RO-EN
Baseline	36.80
Fixed BART	36.29
Tuned BART	**37.96**

Table 8: BLEU scores of the baseline and BART on WMT'16 RO-EN augmented with back-translation data. BART improves over a strong back-translation baseline by using monolingual English pre-training.

Thanks