Pre-Trained Checkpoints for Warm-Starting of Generation Models

Haoyu Song

Sep 24th 2020

A Recent Study On This Topic

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Sascha Rothe
Google Research
rothe@google.com

Shashi Narayan
Google Research
shashinarayan@google.com

Aliaksei Severyn
Google Research
severyn@google.com

Transactions of the Association for Computational Linguistics
Volume 8, 2020
p.264-280

☆ Cited by 27 Related articles

Preliminaries

1. **Autoregressive LM VS Autoencoding LM**

2. **The differences in Pretraining and Finetuning between GPT, BERT**

Preliminaries — Autoregressive LM

$$max_\theta \, log_\theta(x) = \sum_1^T log \, \frac{\exp((h_\theta(x_{1:t-1}))^{\mathrm{T}} e(x_t))}{\sum_{x'} \exp((h_\theta(x_{1:t-1}))^{\mathrm{T}} e(x'))}$$

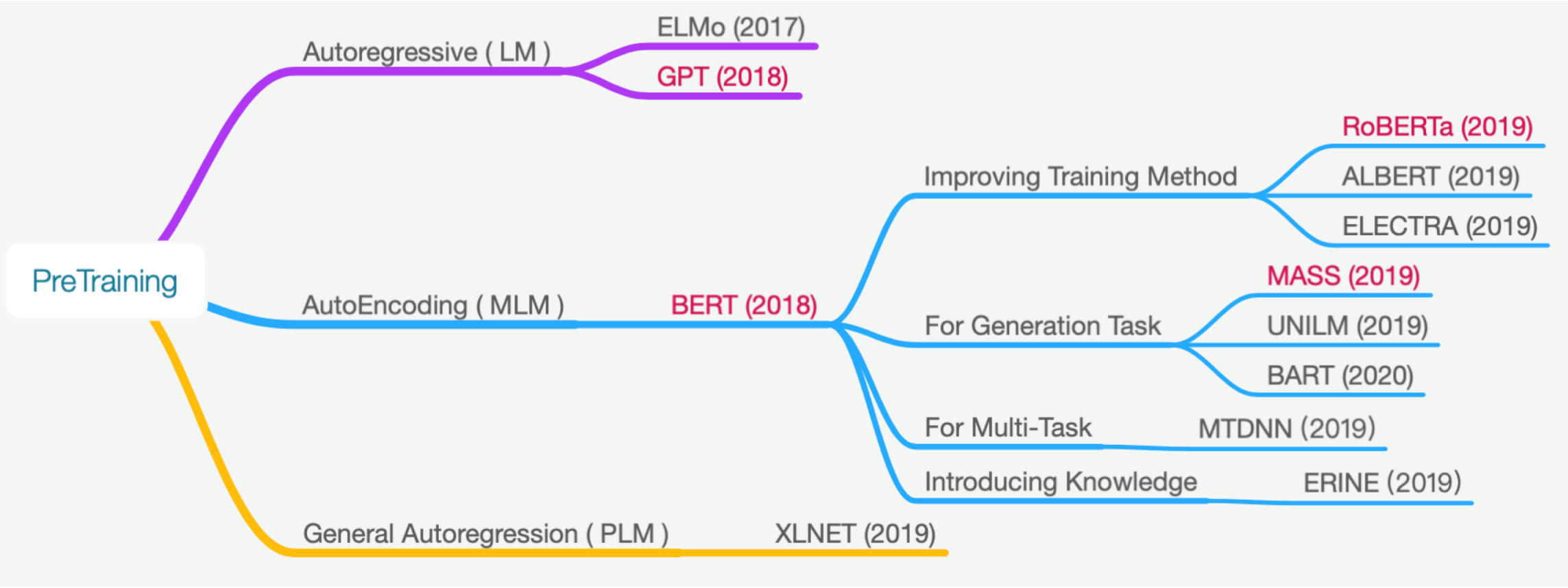
- **Pros:** classical language model, time-tested
- **Cons:** unable to model the context window
- **Examples:** ELMO, GPT 1.0/2.0/3.0

Preliminaries — Autoencoding LM

$$max_\theta \, log_\theta(\bar{x}|\hat{x}) = \sum_1^T mask_t \, log \, \frac{\exp((h_\theta(\hat{x}_t))^{\mathrm{T}} e(x_t))}{\sum_{x'} \exp((h_\theta(\hat{x}_t))^{\mathrm{T}} e(x'))}$$

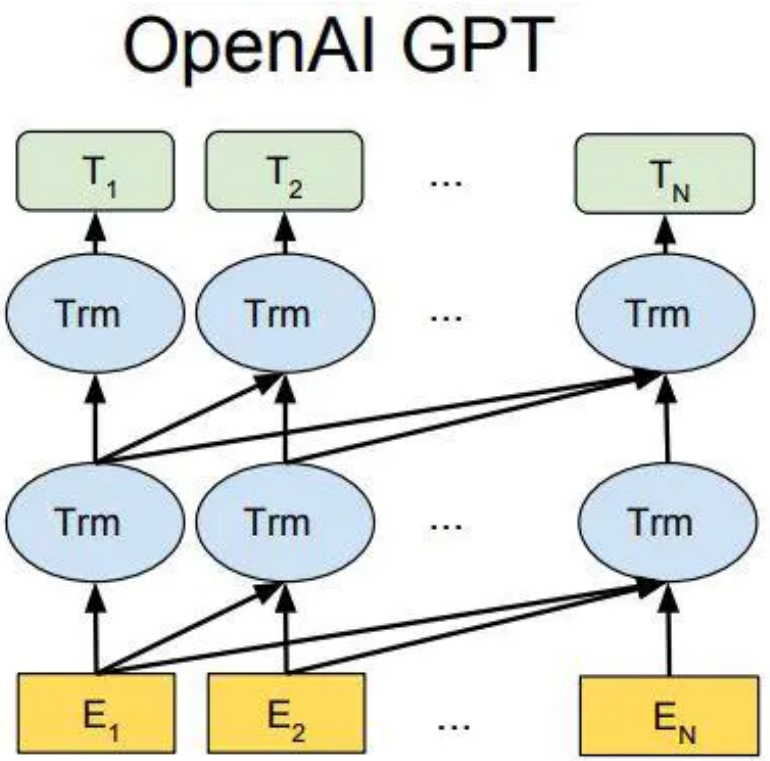
- **Pros:** context sensitive bidirectional representation
- **Cons:** unable to model the context window
- **Examples:** BERT and BERT Variants

A Road Map for Pretrained Models



1. GPT Training

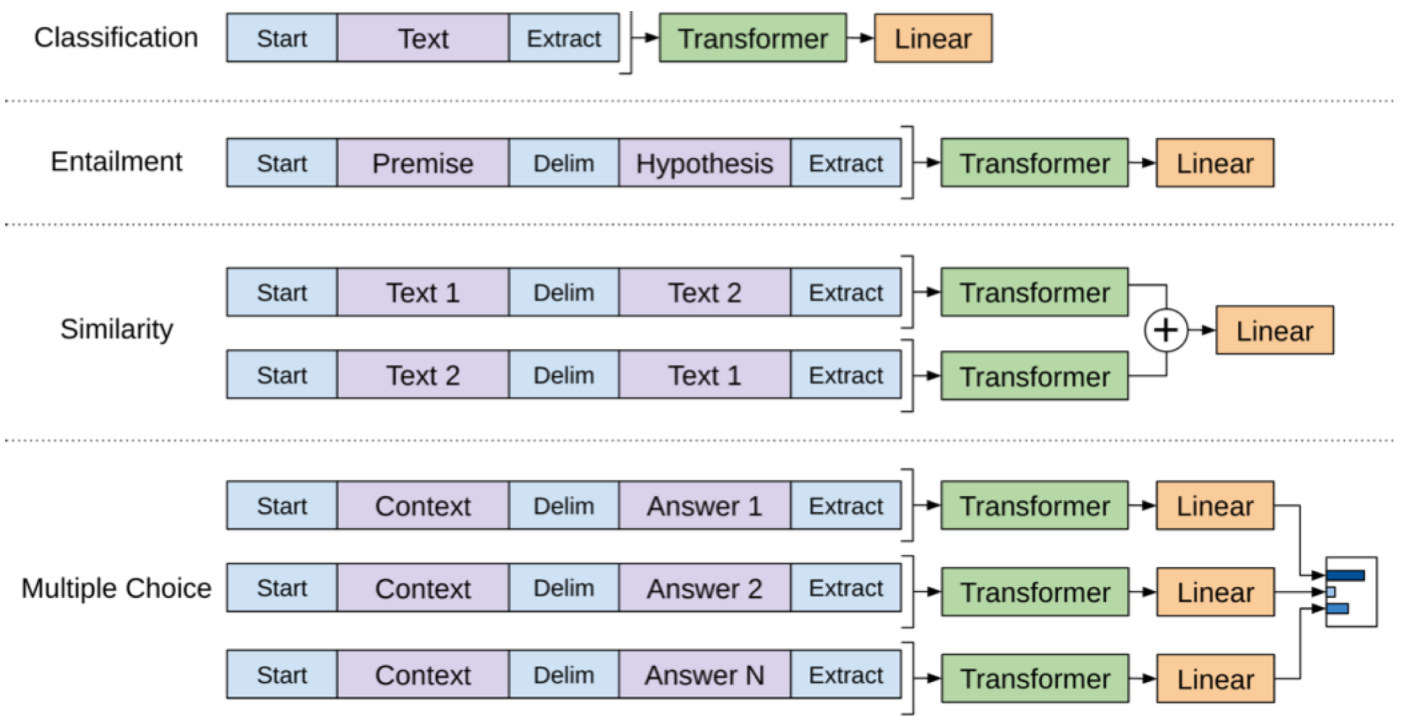
- GPT is a transformer **decoder** in training
  - Word vector of n words in the sentence is added with positional encoding
  - Then input into transformer to output the next word in n+1 position
  - Masked self-attention



Improving Language Understanding by Generative Pre-Training, 2018

1. GPT Finetuning

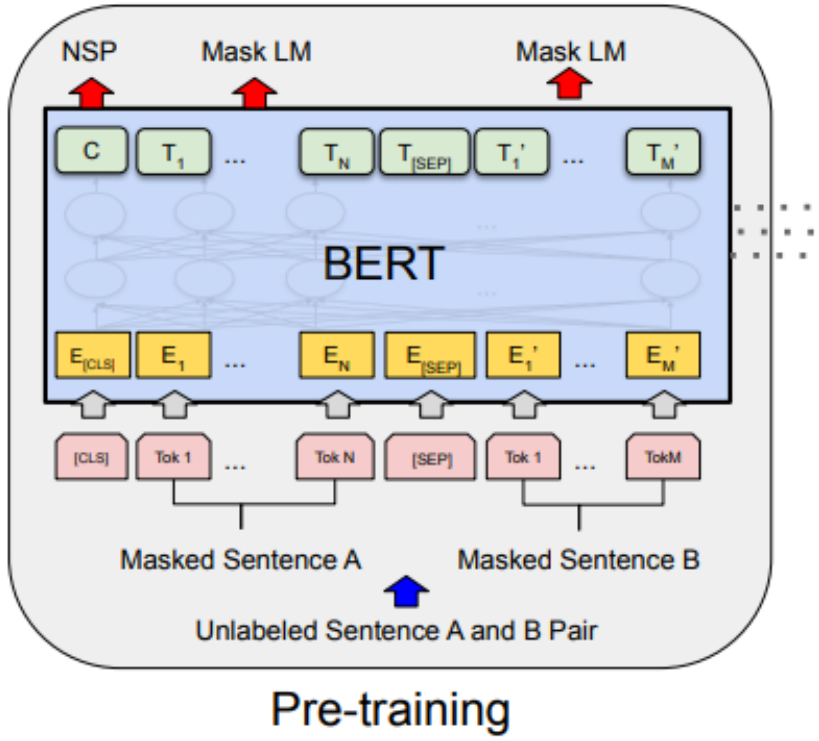
- GPT for NLU needs task-specific modifications



Improving Language Understanding by Generative Pre-Training, 2018

2. BERT Training

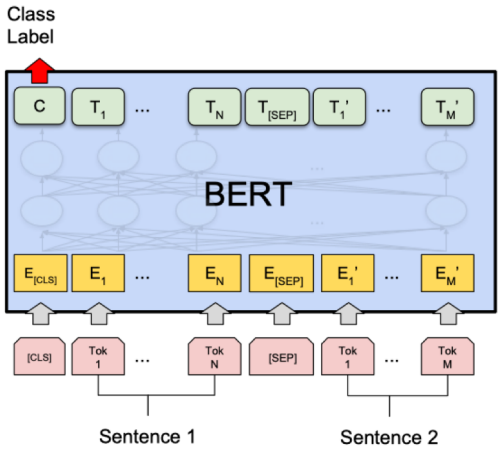
- BERT is a transformer **encoder** in training
  - MLM + NSP
  - Trained to predict the masked word and next sentence
  - self-attention ⟶ embedding ⟶ probability



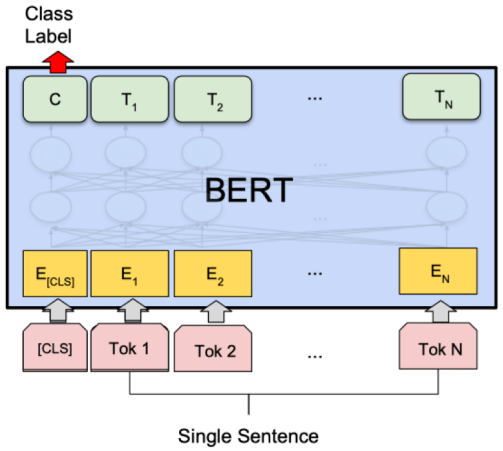
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018

2. BERT Finetuning

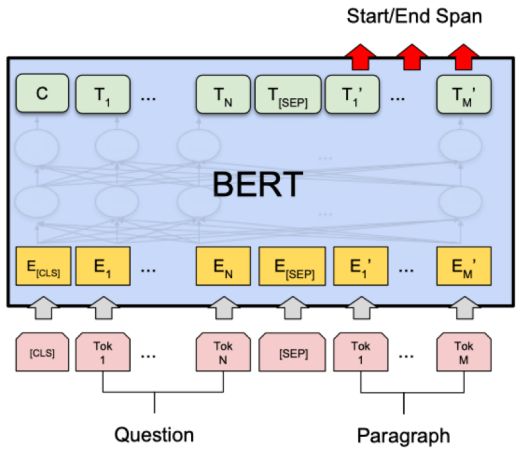
- BERT has unified architecture for all NLU tasks



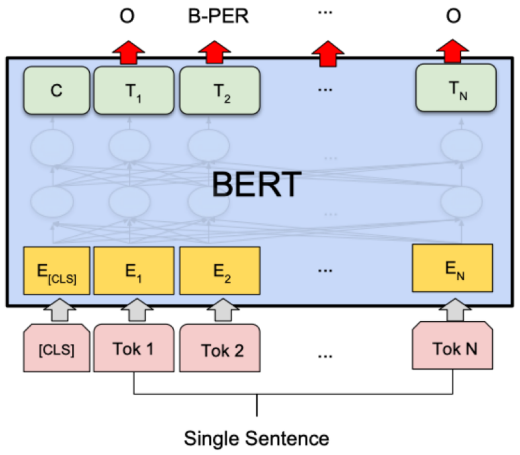
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

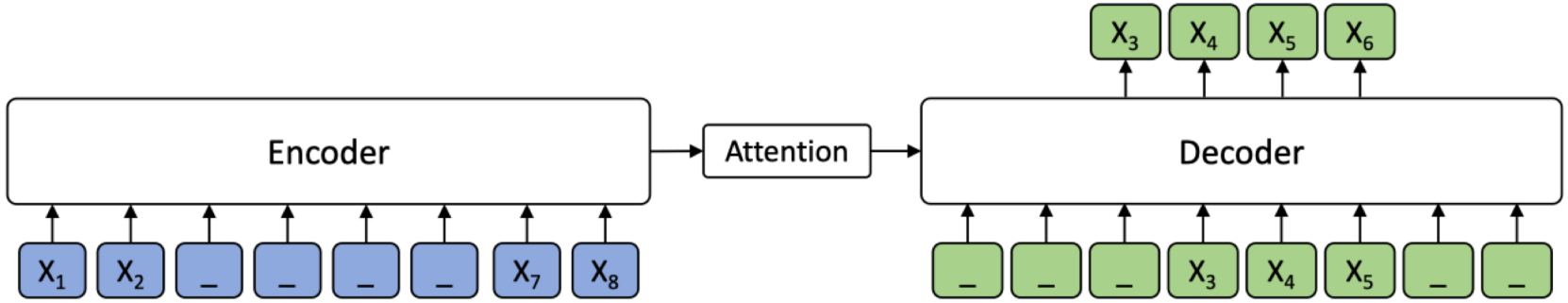
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018

3. RoBERTa Pretraining

- RoBERTa improves the pretraining methods of BERT
  - Dynamic Masking
  - Full-Sentences without NSP
  - Large Mini-Batches
  - Byte-Level BPE

RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019

4. MASS

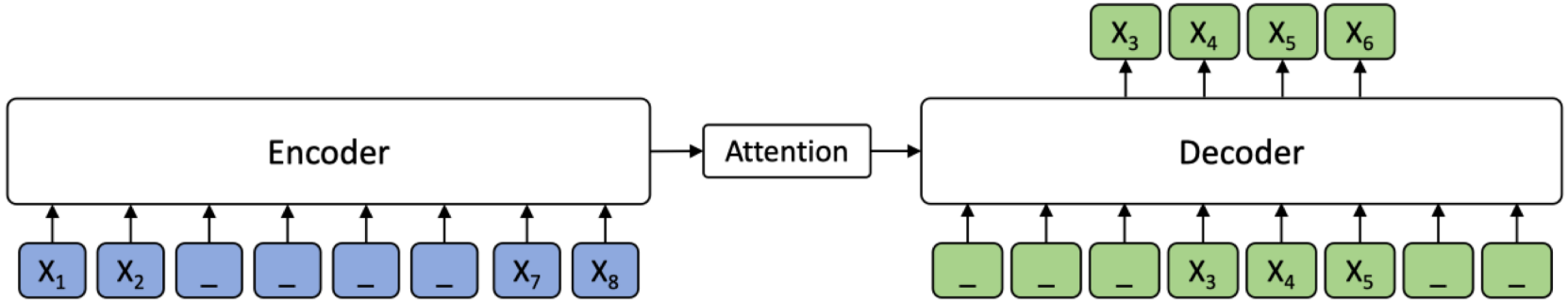


- MASS aims to improve BERT on generation tasks

- "Directly applying a BERT like pre-training method on these natural language generation tasks is not feasible, since BERT is designed for language understanding."

MASS: Masked Sequence to Sequence Pre-training for Language Generation，2019

4. MASS



- MASS aims to improve BERT on generation tasks

- "Directly applying a BERT like pre-training method on these natural language generation tasks is not feasible, since BERT is designed for language understanding."

MASS: Masked Sequence to Sequence Pre-training for Language Generation, 2019

A Question

- Can pre-trained checkpoints, like **BERT**, **RoBERTa**, be leveraged for **sequence generation models** ?

A Question

- Can pre-trained checkpoints, like **BERT**, **RoBERTa**, be leveraged for **sequence generation models** ?
  - They are famous for their performance on NLU tasks, not validated on NLG
    - SST, MNLI, SQuAD, MRPC ···
  - MLM is different from generation process
  - Encoder-only, rather than encoder-decoder

Go Back to Today's Paper

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Sascha Rothe
Google Research
rothe@google.com

Shashi Narayan
Google Research
shashinarayan@google.com

Aliaksei Severyn
Google Research
severyn@google.com

Transactions of the Association for Computational Linguistics
Volume 8, 2020
p.264-280

☆ Cited by 27 Related articles

Background

- Unsupervised pre-training models has revolutionized NLP
- Released checkpoints pushed the SOTA and saving computing time
- The focus has been mainly on the NLU tasks
- This paper demonstrate the efficacy of pre-trained checkpoints for Sequence Generation

BERT, RoBERTa

Basic Architecture

- A Seq2Seq architecture with encoder and decoder both composed of Transformer layers
- **Encoder:** Inherit the BERT Transformer layer implementations
- **Decoder:** Identical to the BERT implementation with two adjustments
  - Self-attention mechanism is masked to look only at the left context
  - Add an encoder-decoder attention mechanism

Basic Hyperparameters

- Most models use base checkpoint
  - 12 layers, hidden size of 768, and 12 attention heads
  - Fine-tuned on the target task using Adam with a learning rate of 0.05

Model Variants

1. **RND2RND**: A Transformer encoder-decoder architecture with all weights initialized randomly

2. **BERT2RND**: A BERT-initialized encoder paired with a randomly initialized decoder. Encoder and decoder share the embedding matrix initialized from a checkpoint.

3. **RND2BERT**: ⋯

4. **BERT2BERT**: A BERT-initialized encoder paired with a BERT-initialized decoder.

5. **BERTSHARE**: Like BERT2BERT, but the parameters between encoder and decoder are shared.

6. **ROBERTASHARE**: Same as BERTSHARE, from RoBERTa checkpoint.

7. **GPT**: A decoder-only architecture with a public GPT-2 checkpoint.

Model Variants

8. **RND2GPT** : …

9. **BERT2GPT**: A BERT-compatible encoder paired with a GPT-2-compatible decoder. BERT vocabulary for the input and the GPT-2 vocabulary for the output.

10. **ROBERTA2GPT**: …

	total	embed.	init.	random
RND2RND	221M	23M	0	221M
BERT2RND	221M	23M	109M	112M
RND2BERT	221M	23M	109M	26M
BERT2BERT	221M	23M	195M	26M
BERTSHARE	136M	23M	109M	26M
ROBERTASHARE	152M	39M	125M	26M
GPT	125M	39M	125M	0
RND2GPT	238M	39M	125M	114M
BERT2GPT	260M	62M	234M	26M
ROBERTA2GPT	276M	78M	250M	26M

Experiment 1: Sentence Fusion

- Sentence Fusion is the problem of combining multiple sentences into a single coherent sentence.

DiscoFuse	100% Exact	100% SARI	10% SARI	1% SARI
(Geva et al., 2019)	51.1	84.5	–	–
Initialized with the base checkpoint (12 layers)				
ROBERTA2GPT	**65.6**	**89.9**	**87.1**	80.3
ROBERTASHARE	65.3	89.7	86.9	**81.2**
BERT2BERT	63.9	89.3	86.1	**81.2**
BERT2RND	63.9	89.3	86.1	80.3
BERTSHARE	63.9	89.2	86.0	80.8
BERT2GPT	61.5	88.4	84.1	70.2
GPT	60.4	88.0	82.9	74.5
RND2BERT	60.0	87.6	82.1	72.8
RND2RND	58.3	86.9	81.5	69.3
RND2GPT	57.6	86.5	81.4	70.6
Initialized with the large checkpoint (24 layers)				
ROBERTASHARE	**66.6**	**90.3**	**87.7**	**81.5**
BERTSHARE	65.3	89.9	86.6	81.4

← Previous SOTA: vanilla transformer with only 7 layers

Training was done for 300k steps with a global batch size of 256.

* **SARI** is a lexical similarity metric which compares the model's output to multiple references and the input in order to assess the model's ability to add, delete and keep an n-gram.

Experiment 2: Split and Rephrase

- The reverse task of sentence fusion, which requires rewriting a long sentence into two or more coherent short sentences

WikiSplit	Exact	SARI	BLEU
(Botha et al., 2018)	14.3	61.5	76.4
Initialized with the base checkpoint (12 layers)			
BERTSHARE	**16.3**	**63.5**	**77.2**
ROBERTASHARE	16.1	63.4	77.1
BERT2BERT	15.6	63.2	77.0
ROBERTA2GPT	15.1	63.2	76.8
BERT2RND	15.9	63.1	76.9
BERT2GPT	14.6	62.4	76.5
RND2BERT	15.2	61.8	76.5
RND2RND	14.6	61.7	76.3
RND2GPT	14.2	61.3	76.2
GPT	14.2	61.1	75.8
Initialized with the large checkpoint (24 layers)			
ROBERTASHARE	16.4	**63.8**	**77.4**
BERTSHARE	**16.6**	63.7	77.3

← Previous SOTA: a bi-directional LSTM with a copy mechanism

Training was done for 300k steps with a global batch size of 256.

* **Exact** is the exact match accuracy

Experiment 3: Machine Translation

- WMT 2014 English-German task, using newstest2014 and newstest2016 evaluation sets
- input and output lengths to 128 tokens
- batch size of 256 and train for 30 epochs
- beam size of 4
- sentence length penalty is set to $\alpha = 0.6$

Experiment 3: Machine Translation

	newstest2014		newstest2016	
	En→De	De→En	En→De	De→En
(Vaswani et al., 2017)	27.3	–	–	–
Transformer (ours)	28.1	31.4	33.5	37.9
KERMIT (Chan et al., 2019)	28.7	31.4	–	–
(Shaw et al., 2018)	29.2	–	–	–
(Edunov et al., 2018)*	**35.0** (33.8)	–	–	–
Initialized with public checkpoints (12 layers) and vocabulary				
Transformer (ours)	23.7	26.6	31.6	35.8
RND2RND	26.0	29.1	32.4	36.7
BERT2RND	**30.1**	**32.7**	34.4	**39.6**
RND2BERT	27.2	30.4	33.2	37.5
BERT2BERT	**30.1**	**32.7**	**34.6**	39.3
BERTSHARE	29.6	32.6	34.4	**39.6**
GPT	16.4	21.5	22.4	27.7
RND2GPT	19.6	23.2	24.2	28.5
BERT2GPT	23.2	31.4	28.1	37.0
Initialized with a custom BERT checkpoint (12 layers) and vocabulary				
BERT2RND	**30.6**	33.5	35.1	**40.2**
BERTSHARE	30.5	**33.6**	**35.5**	40.1
Initialized with a custom BERT checkpoint (24 layers) and vocabulary				
BERT2RND	**31.7**	**34.2**	**35.6**	**41.1**
BERTSHARE	30.5	33.8	35.4	40.9

augment the training set with a massive amount of back-translated sentence pairs

RoBERTa checkpoint is available for English only. So it was excluded in this experiment.

Experiment 3: Machine Translation

Method	Setting	en - fr	fr - en	en - de	de - en	en - ro	ro - en
Artetxe et al. (2017)	2-layer RNN	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	3-layer RNN	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	4-layer Transformer	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	4-layer Transformer	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	6-layer Transformer	33.40	33.30	27.00	34.30	33.30	31.80
MASS	6-layer Transformer	**37.50**	**34.90**	**28.30**	**35.20**	**35.20**	**33.10**
Initialized with public checkpoints (12 layers) and vocabulary							
Transformer (ours)		23.7	26.6	31.6	35.8		
RND2RND		26.0	29.1	32.4	36.7		
BERT2RND		**30.1**	**32.7**	34.4	**39.6**		
RND2BERT		27.2	30.4	33.2	37.5		
BERT2BERT		**30.1**	**32.7**	**34.6**	39.3		
BERTSHARE		29.6	32.6	34.4	**39.6**		
GPT		16.4	21.5	22.4	27.7		
RND2GPT		19.6	23.2	24.2	28.5		
BERT2GPT		23.2	31.4	28.1	37.0		
Initialized with a custom BERT checkpoint (12 layers) and vocabulary							
BERT2RND		**30.6**	33.5	35.1	**40.2**		
BERTSHARE		30.5	**33.6**	**35.5**	40.1		
Initialized with a custom BERT checkpoint (24 layers) and vocabulary							
BERT2RND		**31.7**	**34.2**	**35.6**	**41.1**		
BERTSHARE		30.5	33.8	35.4	40.9		

A comparable result between MASS and this paper is on the BLEU of newstest2016 en-de & de-en

Experiment 4: Abstractive Summarization

	Gigaword			CNN/Dailymail			BBC XSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Lead	–	–	–	39.60	17.70	36.20	16.30	1.61	11.95
PtGen	–	–	–	39.53	17.28	36.38	29.70	9.21	23.24
ConvS2S	35.88	17.48	33.29	–	–	–	31.89	11.54	25.75
MMN	–	–	–	–	–	–	32.00	12.10	26.00
Bottom-Up	–	–	–	41.22	18.68	38.34	–	–	–
MASS	*38.73*	19.71	*35.96*	–	–	–	–	–	–
TransLM	–	–	–	39.65	17.74	36.85	–	–	–
UniLM	–	–	–	*43.47*	*20.30*	*40.63*	–	–	–
Initialized with the base checkpoint (12 layers)									
RND2RND	36.94	18.71	34.45	35.77	14.00	32.96	30.90	10.23	24.24
BERT2RND	37.71	19.26	35.26	38.74	17.76	35.95	38.42	15.83	30.80
RND2BERT	37.01	18.91	34.51	36.65	15.55	33.97	32.44	11.52	25.65
BERT2BERT	38.01	19.68	35.58	39.02	17.84	36.29	37.53	15.24	30.05
BERTSHARE	38.13	***19.81***	**35.62**	39.09	18.10	36.33	38.52	16.12	31.13
ROBERTASHARE	**38.21**	19.70	35.44	**40.10**	**18.95**	**37.39**	**39.87**	**17.50**	**32.37**
GPT	36.04	18.44	33.67	37.26	15.83	34.47	22.21	4.89	16.69
RND2GPT	36.21	18.39	33.83	32.08	8.81	29.03	28.48	8.77	22.30
BERT2GPT	36.77	18.23	34.24	25.20	4.96	22.99	27.79	8.37	21.91
ROBERTA2GPT	37.94	19.21	35.42	36.35	14.72	33.79	19.91	5.20	15.88
Initialized with the large checkpoint (24 layers)									
BERTSHARE	38.35	19.80	35.66	39.83	17.69	37.01	38.93	16.35	31.52
ROBERTASHARE	**38.62**	19.78	**35.94**	**40.31**	18.91	**37.62**	***41.45***	***18.79***	***33.90***

← Non-Pre-trained Models

← Pre-trained Models

	Length	Repetitions	Quality
RND2RND	20.90	29.76	-0.103
RND2GPT	21.49	**16.28**	-0.303
BERTSHARE	20.71	27.03	-0.097
ROBERTASHARE	**21.70**	28.68	**0.153**
GOLD	24.61	4.66	0.347

Table 7: Qualitative and human evaluations of BBC ex treme summaries. The lowest numbers for repetition and the highest numbers for quality are bold faced. Se the text for details.

Some Ablation Studies

- Combining BERT and GPT-2 into a single model did not work and often underperformed than a randomly initialized baseline.

- **GPT-2 checkpoint performed relatively poorly.** Against the intuition that GPT-2 initialized decoders will be strong natural language generators.

- **Initializing only Layers (excluding word embedding)**: Both GPT and BERT are way behind the fully initialized setup.

- **Initializing only Embeddings**: almost no improvement over the fully randomly initialized model

- **Initializing a Subset of Layers**: results from best layer is still outperformed by the base BERT

Discussion

- This paper shows the way to leverage publicly available pre-trained checkpoints for warm-starting sequence generation models

Some Open Questions:

- Do pre-trained models need to be modified for generation tasks?

- Should future works follow the private BART or just use the publicly available checkpoints?

Reference

1. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks
2. RoBERTa: A Robustly Optimized BERT Pretraining Approach
3. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
4. Improving Language Understanding by Generative Pre-Training
5. MASS: Masked Sequence to Sequence Pre-training for Language Generation
6. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

END
Q & A