Unified Language Model Pre-training for Natural Language Understanding and Generation

- Li Dong et al. @ Microsoft Research [1]

Presenter - Anup Deshmukh, 20837751
Advised by Prof. Ming Li

What is Text Summarization?

- The goal here is to condense a document into a shorter version while preserving most of its meaning.
- Abstractive Summarization - Generate summaries containing novel words and phrases not featured in the source text. **(Sequence to sequence problem)**
- Extractive Summarization - Identifying and subsequently concatenating the most important sentences in a document. **(Binary classification problem)**

How to evaluate summarization tasks?

- ROUGE score: Recall-Oriented Understudy for Gisting Evaluation
  - ROUGE-N: Overlap of N-grams between the system and reference summaries
  - ROUGE-2$_{recall}$: (number of overlapping bigrams) / (number of bigrams in the reference summary)
  - ROUGE-2$_{precision}$: (number of overlapping bigrams) / (number of bigrams in the system summary)

System Summary :

```
the cat was found under the bed
```

Reference Summary :

```
the cat was under the bed
```

System Summary Bigrams:

```
the cat,
cat was,
was found,
found under,
under the,
the bed
```

Reference Summary Bigrams:

```
the cat,
cat was,
was under,
under the,
the bed
```

How to evaluate summarization tasks?

- ROUGE score: Recall-Oriented Understudy for Gisting Evaluation
  - ROUGE-L: Longest common subsequence (LCS)

System Summary :

```
the cat was found under the bed
```

Reference Summary :

```
the cat was under the bed
```

System Summary Bigrams:

```
the cat,
cat was,
was found,
found under,
under the,
the bed
```

Reference Summary Bigrams:

```
the cat,
cat was,
was under,
under the,
the bed
```

Motivation

- BERT: Although BERT significantly improves the performance of a wide range of natural language understanding tasks, its bidirectionality nature makes it difficult to be applied to natural language generation tasks
- UniLM: Multi-layer Transformer network, jointly pre-trained on large amounts of text, optimized for three types of unsupervised language modeling objectives.
  - But unlike BERT which is used mainly for NLU tasks, UniLM can be configured, using different self-attention masks, to aggregate context for different types of language models, and thus can be used for both NLU and NLG tasks.

Motivation

	ELMo	GPT	BERT	UniLM
Left-to-Right LM	□	□		□
Right-to-Left LM	□		□	□
Bidirectional LM				□
Sequence-to-Sequence LM				□

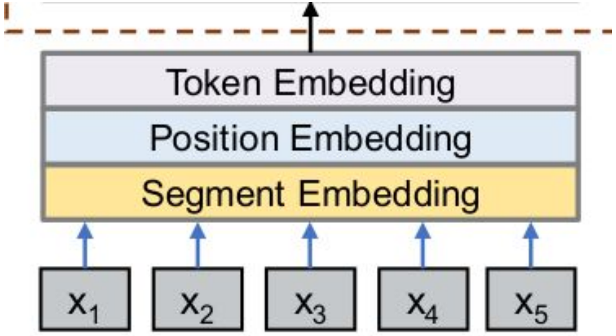
- For a **sequence-to-sequence LM**, the context of the to-be-predicted word in the second (target) sequence consists of all the words in the first (source) sequence and the words on the its left in the target sequence

UniLM (Unified pre-trained Language Model)

- UniLM:
    - Pre-trained using three types of language models: unidirectional, bidirectional, and sequence to sequence prediction.
    - Employs a shared Transformer network and utilizes specific self attention masks
    - It achieves new state of the art results on 5 natural language generation tasks

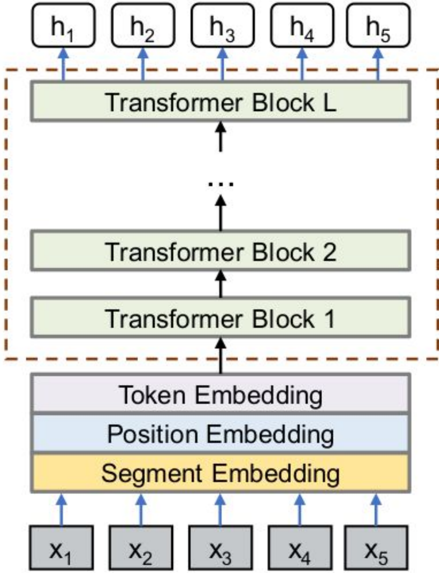
UniLM (Unified pre-trained Language Model)

- Input representation
  - Follows that of BERT
  - special start-of-sequence ([SOS]) token at the beginning of input
  - special end-of-sequence ([EOS]) token at the end of each segment.



UniLM (Unified pre-trained Language Model)

- Backbone Network: Multi-Layer Transformer
  - The idea: In order to control the access to the context of the word token to be predicted, authors employ different masks for self-attention.



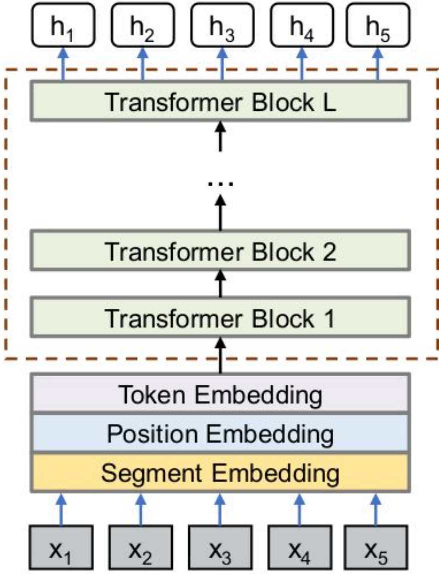
UniLM (Unified pre-trained Language Model)

- Backbone Network: Multi-Layer Transformer

$$\mathbf{Q} = \mathbf{H}^{l-1}\mathbf{W}_l^Q, \quad \mathbf{K} = \mathbf{H}^{l-1}\mathbf{W}_l^K, \quad \mathbf{V} = \mathbf{H}^{l-1}\mathbf{W}_l^V$$

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases}$$

$$\mathbf{A}_l = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V}_l$$



UniLM (Unified pre-trained Language Model)

- Pre-training Objectives
  - Randomly choose some tokens in the input and replace them with the special token, [MASK]
  - Then, feed their corresponding output vectors computed by the Transformer network into a softmax classifier to predict the masked token.

UniLM (Unified pre-trained Language Model)

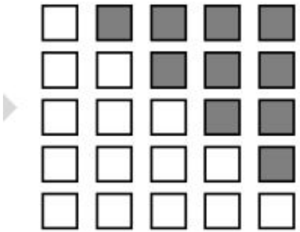
- Pre-training Objectives
  - The token masking probability is 15%
    - Among masked positions, 80% of the time we replace the tokens with [MASK]
    - 10% of the time with the random token
    - Keeping the original token for the rest 10% of the time

UniLM (Unified pre-trained Language Model)

- Pre-training Objectives
    - If we used [MASK] 100% of the time the model wouldn't necessarily produce good token representations for non-masked words. The non-masked tokens were still used for context, but the model was optimized for predicting masked words.
    - If we used [MASK] 90% of the time and random words 10% of the time, this would teach the model that the observed word is *never* correct.

UniLM (Unified pre-trained Language Model)

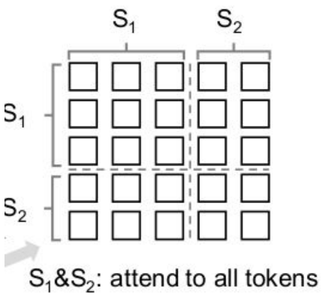
- Pre-training Objectives
  - The overall training objective the sum of different types of LM objectives.
- Unidirectional LM
  - Example: x_1 x_2 [MASK] x_4
  - Using a triangular matrix for the self-attention mask M



S_1: attend to left context

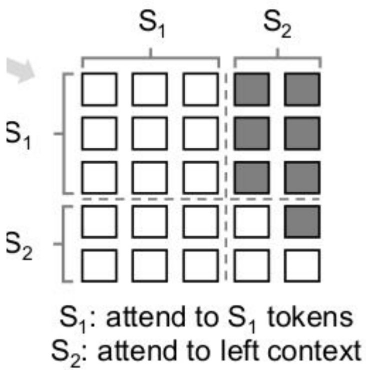
UniLM (Unified pre-trained Language Model)

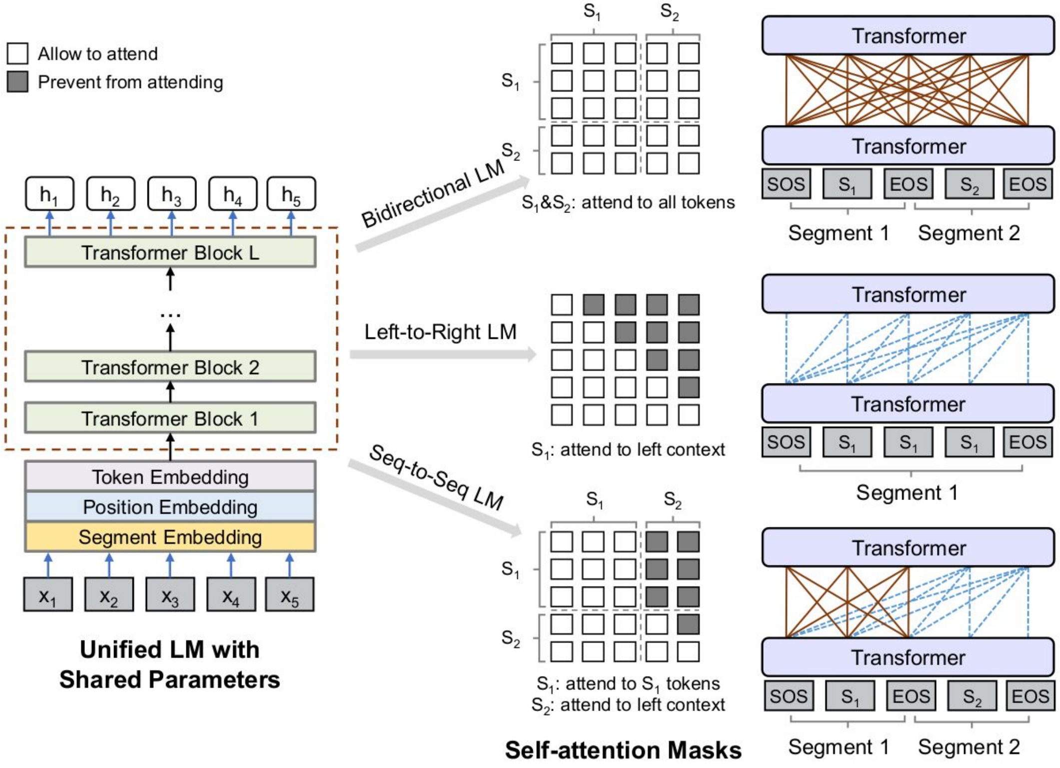
- Bidirectional LM
    - bidirectional LM allows all tokens to attend to each other in prediction
    - It encodes contextual information from both directions
    - The self-attention mask M is a zero matrix, so that every token is allowed to attend across all positions in the input sequence



UniLM (Unified pre-trained Language Model)

- Sequence-to-Sequence LM
  - Example: [SOS] t_1 t_2 [EOS] t_3 t_4 t_5 [EOS]
  - both t_1 and t_2 have access to the first four tokens
  - t_4 can only attend to the first six tokens.





Allow to attend
Prevent from attending
h1 h2 h3 h4 h5
Transformer Block L
...
Transformer Block 2
Transformer Block 1
Token Embedding
Position Embedding
Segment Embedding
x1 x2 x3 x4 x5
Unified LM with Shared Parameters
Bidirectional LM
Left-to-Right LM
Seq-to-Seq LM
S1 S2
S1&S2: attend to all tokens
S1: attend to left context
S1: attend to S1 tokens
S2: attend to left context
Self-attention Masks
Transformer
SOS S1 EOS S2 EOS
Segment 1 Segment 2
SOS S1 S1 S1 EOS
Segment 1

Backbone Network	LM Objectives of Unified Pre-training	What Unified LM Learns	Example Downstream Tasks
Transformer with shared parameters for all LM objectives	Bidirectional LM	Bidirectional encoding	GLUE benchmark Extractive question answering
	Unidirectional LM	Unidirectional decoding	Long text generation
	Sequence-to-Sequence LM	Unidirectional decoding conditioned on bidirectional encoding	Abstractive summarization Question generation Generative question answering

UniLM (Unified pre-trained Language Model)

- Fine tuning for text summarization
  - Let S1 and S2 denote source and target sequences, respectively
  - [SOS]   S1   [EOS]   S2   [EOS]
  - The model is fine-tuned by masking some percentage of tokens in the target sequence at random, and learning to recover the masked words.
  - The training objective is to maximize the likelihood of masked tokens given context

UniLM (Unified pre-trained Language Model)

- Fine tuning for text summarization (further details)
  - CNN/DailyMail and Gigaword datasets are used for model fine tuning and evaluation
  - Authors fine-tune our model on the training set for 30 epochs
  - The masking probability is 0.7
  - During decoding, we use beam search with beam size of 5

UniLM (Unified pre-trained Language Model)

- Experiments and results (CNN/DailyMail abstractive summarization)

	RG-1	RG-2	RG-L
Extractive Summarization			
LEAD-3	40.42	17.62	36.67
Best Extractive [27]	43.25	**20.24**	39.63
Abstractive Summarization			
PGNet [37]	39.53	17.28	37.98
Bottom-Up [16]	41.22	18.68	38.34
S2S-ELMo [13]	41.56	18.94	38.47
UniLM	**43.33**	20.21	**40.51**

UniLM (Unified pre-trained Language Model)

- Experiments and results (Gigaword abstractive summarization)

	RG-1	RG-2	RG-L
10K Training Examples			
Transformer [43]	10.97	2.23	10.42
MASS [39]	25.03	9.48	23.48
UniLM	**32.96**	**14.68**	**30.56**
Full Training Set			
OpenNMT [23]	36.73	17.86	33.68
Re3Sum [4]	37.04	19.03	34.46
MASS [39]	37.66	18.53	34.89
UniLM	**38.45**	**19.45**	**35.75**

UniLM (Unified pre-trained Language Model)

- The UniLM has three main advantages
  - First, the unified pre-training procedure leads to a single Transformer LM that uses the shared parameters and architecture for different types of LMs
  - No overfitting to any single LM task
  - In addition to its application to NLU tasks, the use of UniLM as a sequence-to-sequence LM, makes it a natural choice for NLG, such as abstractive summarization and question generation

References

- [1] Dong, Li, et al. "Unified language model pre-training for natural language understanding and generation." *Advances in Neural Information Processing Systems*. 2019.
- [2] Liu, Yang, and Mirella Lapata. "Text summarization with pretrained encoders." *arXiv preprint arXiv:1908.08345* (2019).
- [3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [4] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Any Questions?

Thank you.