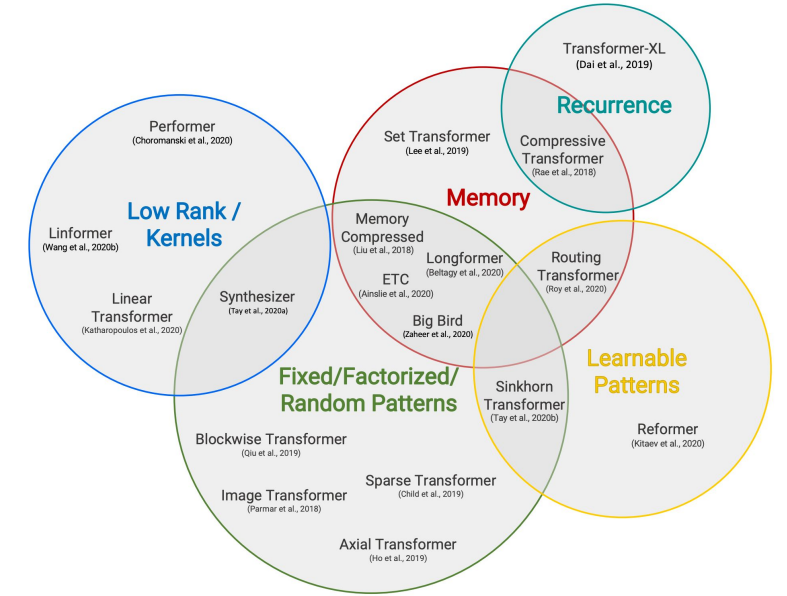
Conclusion and future work

Conclusion

- Long sequence NLP presents many challenges for current models.
  - Long range dependencies often requires complex reasoning and forces models to both locate relevant information and combine it.
  - Popular methods like Transformers need algorithmic modifications to support long sequences.
- Recent advances in efficient transformers open the door to long sequence NLP, and show promising improvements over shorter baselines while being easier to use.
- We can adapt and re-use shorter models to bootstrap training longer ones.

UWNLP AI2

Future work - long transformers evaluation



Many recent papers propose improvements for long sequence transformers, but most of them are not evaluated on NLP benchmarks.

How well do improvements on general benchmarks transfer to NLP?

Fig reference: Tay et al., 2020

Future work - pretraining objectives

Current pretraining objectives only require local context, in most cases.

What additional pretraining objectives are better suited for long documents?

Future work - long output sequences

Current encoder-decoder approaches to long NLP assume the output sequence is short, and apply global attention across the entire decoder.

How can we extend current methods to handle long output sequences in addition to long input sequences?

UWNLP AI2

Future work - retrieval vs. long sequence

The ability to process long sequences blurs the distinction between two-stage retrieval methods and a single stage with long input.

How to best balance retrieval components and long sequence processing (especially in open domain setting)?

UWNLP AI2

Future work - large models

As transformers increase in size, the compute due to self-attention component reduces relative to feed-forward layers.

```
Compute = 2 x n_layers x hidden_size (12 x hidden_size + seqlen)
```
[1]

For large models, `12 x hidden_size >> seqlen`

So impact of sequence length on the total compute is minimal

Do very large transformers require re-thinking long sequence scaling?

[1] Kaplan et. al., 2020

UWNLP AI2

Please join us for live Q&A (check tutorial schedule)

Thank you!