Not all parameters are born equal! Attention is mostly what you need!

Nikolay Bogoychev
n.bogoych@ed.ac.uk

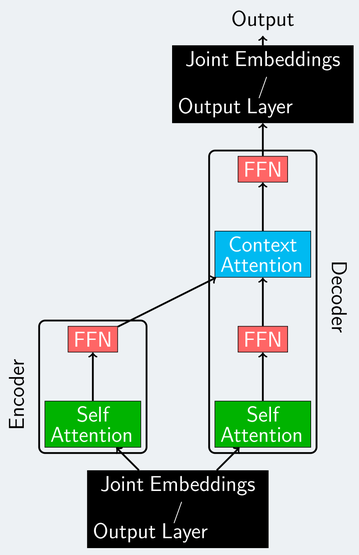
Introduction

- Transformers achieve remarkable results on a variety of tasks
- This is due to the extremely large number of parameters
- And subject to backpropagation and availability of GPU resource.

What if we don't backpropagate through all parameters?
Are some parameters more important than others?

Frozen and Random Transformer components

A simplified illustration of a transformer.



- A transformer has 3 components: Embeddings, Attention and a Feed-Forward Neural network layer.
- We experiment with initialising one or more of them to frozen and random.
- We measure the impact on the model quality.

Transformer-big experiments

WMT18 Turkish-English with frozen and random components.
213M parameters

	Component EMB	ATT	FFN	BLEU	Epochs	Parameter ratio Trainable/All
(0)	✓	✓	✓	24.3	19	1
One frozen component						
(1)	✗	✓	✓	22.6	26	.82
(2)	✓	✗	✓	22.3	23	.64
(3)	✓	✓	✗	23.2	26	.52
Diagonal zeroed frozen component						
(2.1)	✓	✗	✓	19.4	24	.64
(3.1)	✓	✓	✗	22.9	20	.52
Multiple frozen components						
(4)	✗	✓	✗	21.5	36	.35
(5)	✗	✗	✓	20.8	37	.47
(6)	✓	✗	✗	4.4	25	.17

- Attention and FFN have similar importance for the model.
- Embeddings seems to provide somewhat complementary information.
- Over 80% of the performance can be retained with just 35% of parameters.
- Random components are much more useful than diagonal-zero'd components. **The trainable components make use of the available random transformation.**

Reducing FFN width

Reducing the width of the FFN layer to 1024 from 4096.
137M parameters

	Component EMB	ATT	FFN	BLEU	Epochs	Parameter ratio Trainable/All
trans-big baseline				24.3	19	1
(0)	✓	✓	✓	23.2	22	1
One frozen component						
(1)	✗	✓	✓	22.0	38	.73
(2)	✓	✗	✓	20.1	36	.55
(3)	✓	✓	✗	23.0	18	.82
Multiple frozen components						
(4)	✗	✓	✗	21.6	34	.45
(5)	✗	✗	✓	18.0	98	.18

- Smaller models take more epochs to converge and are also more sensitive to component freezing.
- System (0) and (3) are achieve nearly the same quality. Training the small FFN doesn't make a big difference.
- The width of a component is more important than whether it's trainable or not.

LM

A transformer language model trained on 78M sentences.
38M parameters

	Component EMB	ATT	FFN	PPL	Epochs	Parameter ratio Trainable/All
(0)	✓	✓	✓	37.4	6	1
One frozen component						
(1)	✗	✓	✓	118.4	6	.53
(2)	✓	✗	✓	47.5	6	.81
(3)	✓	✓	✗	50.3	6	.64
Multiple frozen components						
(4)	✗	✓	✗	209.3	6	.18
(5)	✗	✗	✓	157.3	6	.36
(6)	✓	✗	✗	131.7	6	.46

- Embeddings much more important than Attenion or FFN unlike translation model experiments.
- Much larger drop in quality compared to translation experiments when freezing components.
- Results likely task specific.

Implications and Conclusion

- A small subset of the big neural network by itself is achieves surprisingly good performance.
- Random components are surprisingly good. Their size is more important than whether they are trainable or not.
- Do we really need high quality pretrained embeddings to use for downstream tasks, if random ones are nearly as good? Questions for pretrained-your-sesame-street-character models.
- Can we do compact neural networks with on-the-fly generated parameters during inference?