Chatbot

- **Video:** Tasks with Long Sequences — 2 min
- **Reading:** Tasks with Long Sequences — 10 min
- **Reading:** Optional AI Storytelling — 15 min
- **Video:** Transformer Complexity — 3 min
- **Reading:** Transformer Complexity — 10 min
- **Video:** LSH Attention — 4 min
- **Reading:** LSH Attention — 10 min
- **Reading:** Optional KNN & LSH Review — 20 min
- **Lab:** Ungraded Lab: Reformer LSH — 1h
- **Video:** Motivation for Reversible Layers: Memory! — 2 min
- **Reading:** Motivation for Reversible Layers: Memory! — 10 min
- **Video:** Reversible Residual Layers — 5 min
- **Reading:** Reversible Residual Layers — 10 min
- **Lab:** Ungraded Lab: Revnet — 1h
- **Video:** Reformer — 2 min
- **Reading:** Reformer — 10 min
- **Reading:** Optional Transformers beyond NLP — 20 min
- **Reading:** Acknowledgments — 10 min

Heroes of NLP: Quoc Le

Assignment

Course Resources

Transformer Complexity

One of the biggest issues with the transformers is that it takes time and a lot of memory when training. Concretely here are the numbers. If you have a sequence of length L, then:

L=100	L^2 = 10K	(0.001s at 10M ops/s)
L=1000	L^2 = 1M	(0.1s at 10M ops/s)
L=10000	L^2 = 100M	(10s at 10M ops/s)
L=100000	L^2 = 10B	(1000s at 10M ops/s)

So if you have N layers, that means your model will take N times more time to complete. As L gets larger, the time quickly increases.

- Attention: softmax(QK^T)V
- Q, K, V are all [L, d_model]
- QK^T is [L, L]
- Save compute by using area of interest for large L

When you are handling long sequences, you usually don't need to consider all L positions. You can just focus on an area of interest instead. For example, when translating a long text from one language to another, you don't need to consider every word at once. You can instead focus on a single word being translated, and those immediately around it, by using attention.

To overcome the memory requirements you can recompute the activations. As long as you do it efficiently, you will be able to save a good amount of time and memory. You will learn this week how to do it. Instead of storing N layers, you will be able to recompute them when doing the back-propagation. That combined with local attention, will give you a much faster model that works at the same level as the transformer you learned about last week.

Mark as completed