

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

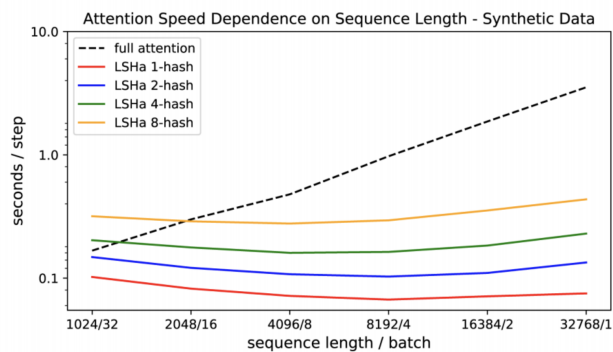
Course Resources



Reformer

The reformer allows you to fit up to 1 million tokens on a single 16 gigabyte GPU. It is designed to handle context windows of up to 1 million words. It combines two techniques to solve the problems of attention and memory allocation which are bottlenecks for the transformer networks.

Reformer uses locality sensitive hashing, which you saw earlier in this specialization, to reduce the complexity of attending over long sequences. It also uses reversible residual layers to more efficiently use the memory available. In the picture below you can see how the reformer performs when compared to a normal full-attention model.

[Mark as completed](#)