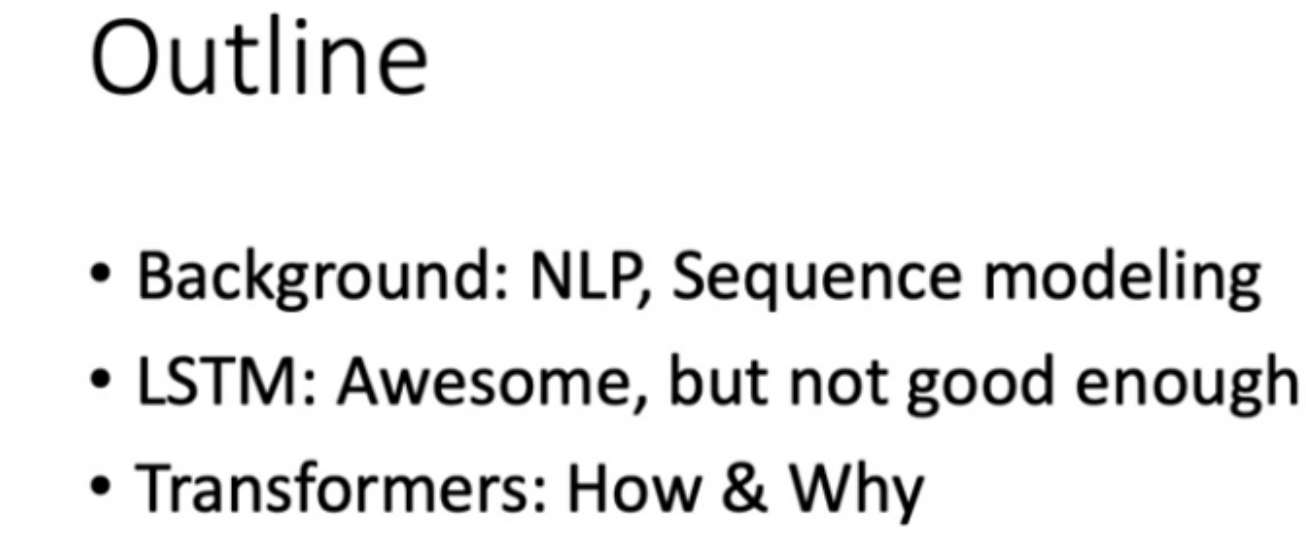
Source

LSTM is dead. Long Live Transformers!

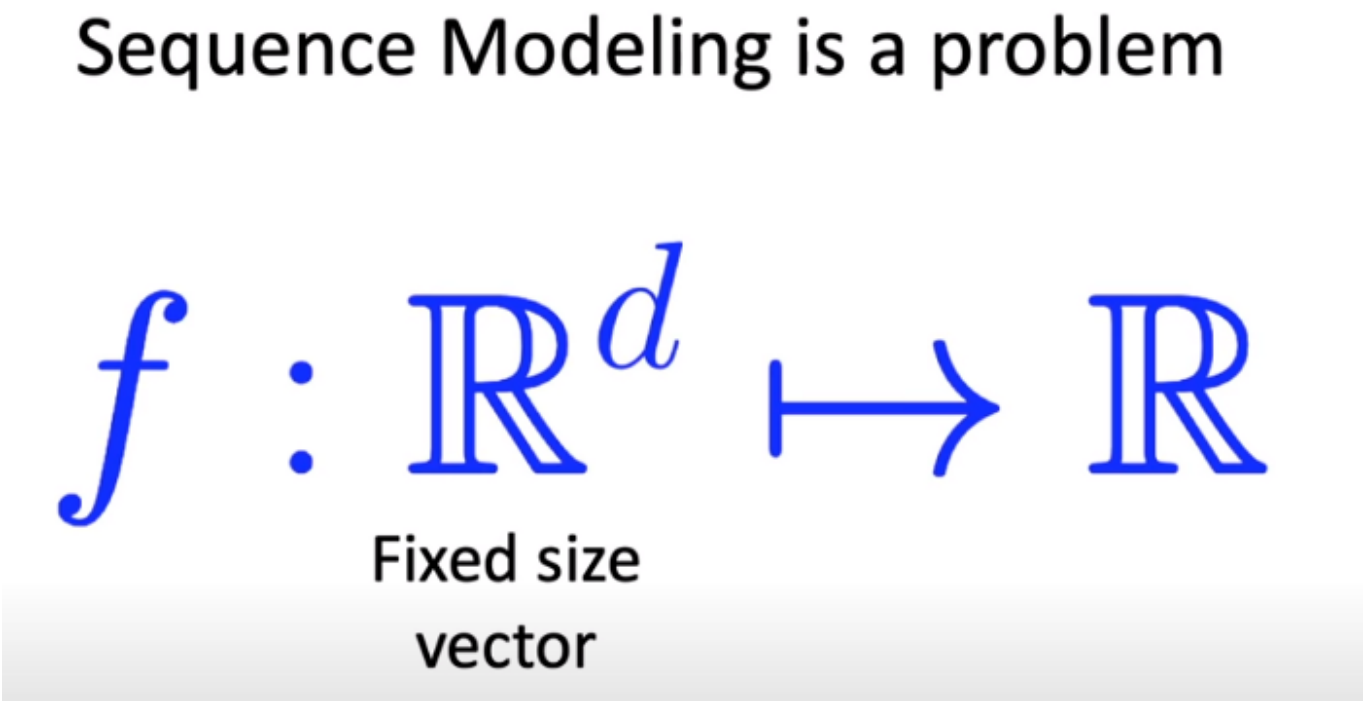
Jae Duk Seo Follow
Jun 7 · 5 min read ★

Why is LSTM Dead? What is the change happening right now that is KILLING LSTM?

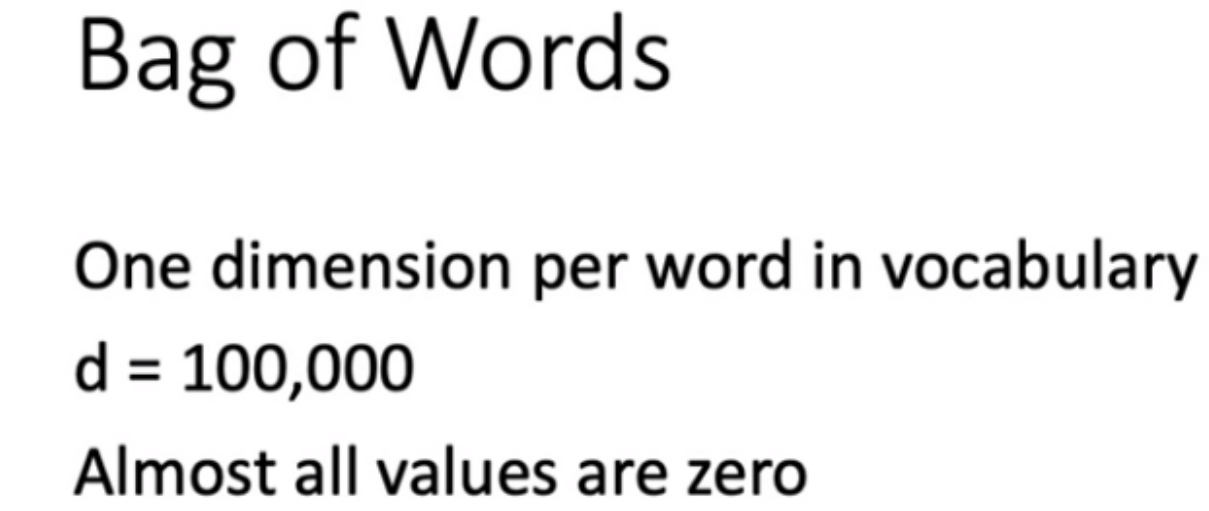


Why LSTM is awesome but why it is not enough, and why attention is making a huge impact.

LSTM has a hard time understanding the full document, how can the model understand everything. It is a long document, how can we make this document to a fixed-sized vector?

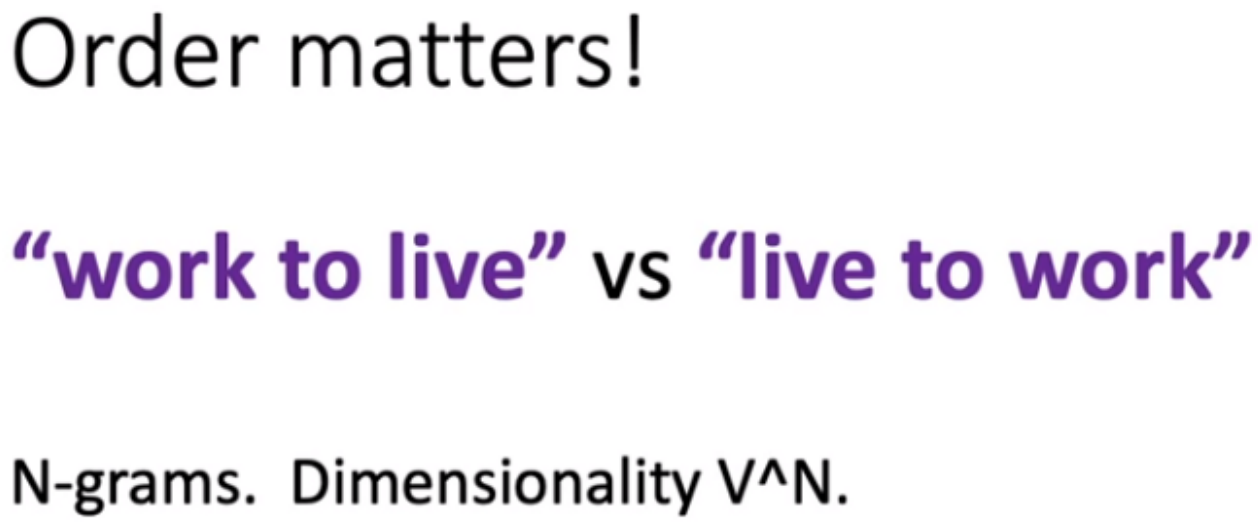


This is hard since the document is not a fixed size length, the classic way to do this is to use Bag of Words.

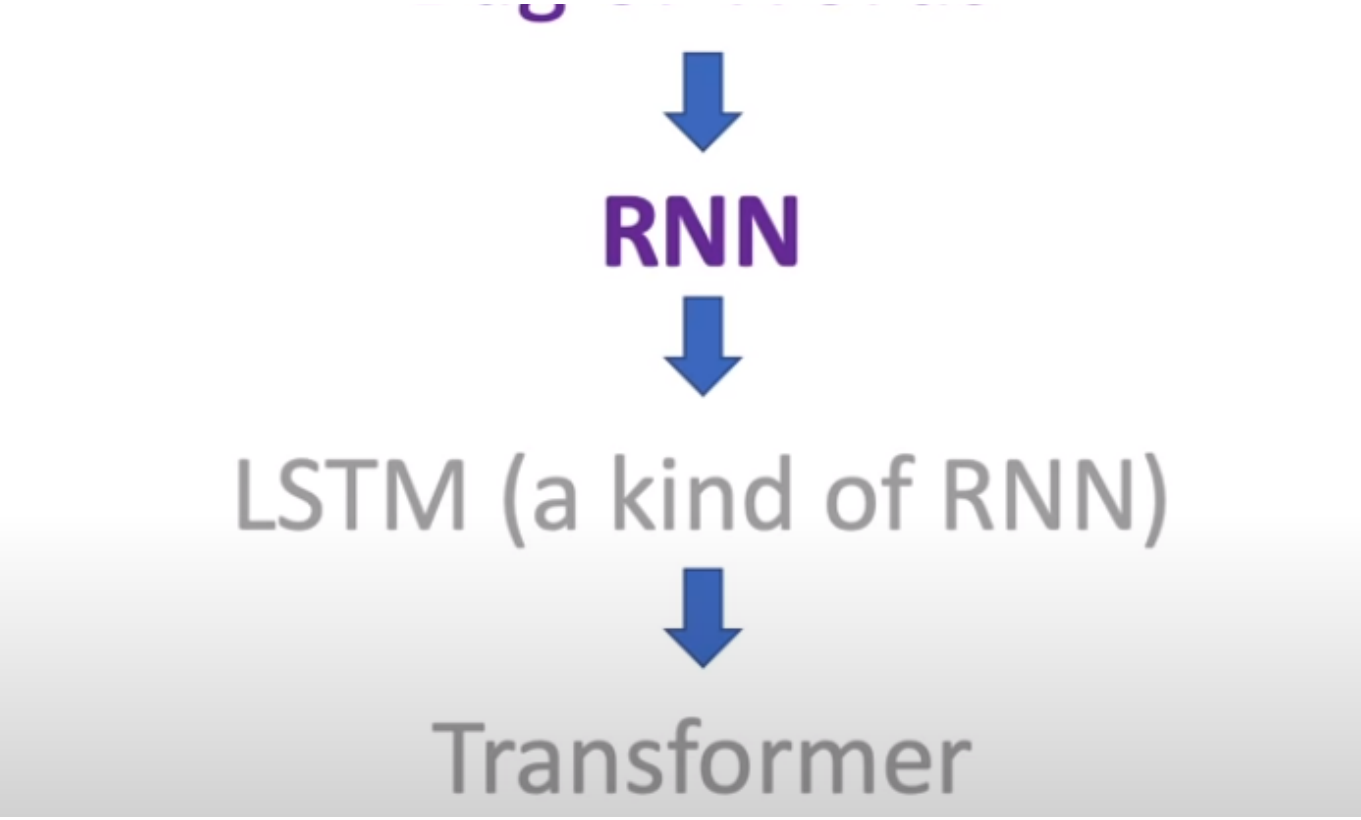


There is a lot of waste since most of the values are going to be zero, indicating that the document does not use a certain word. There would be a better way to encode all of this information, not that effective.

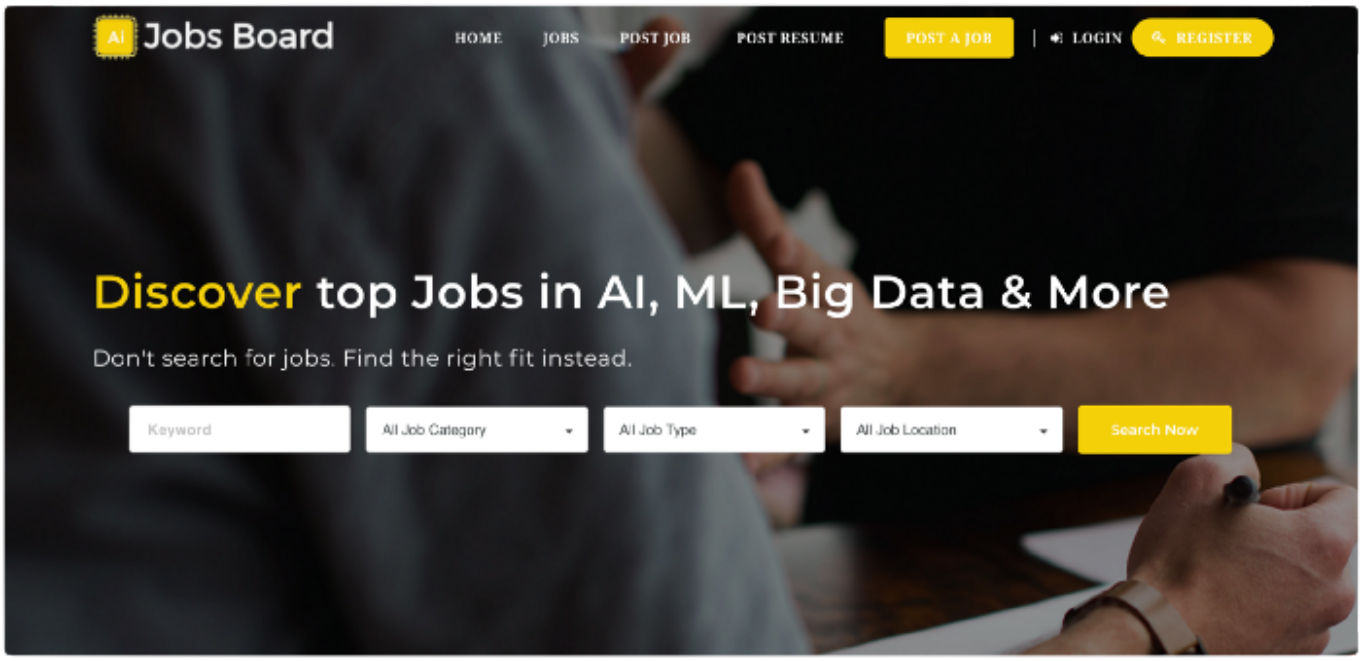
It does work, and the order really matters.



A more serious solution is to use N-grams, but the thing is it gets really high dimensional vector.

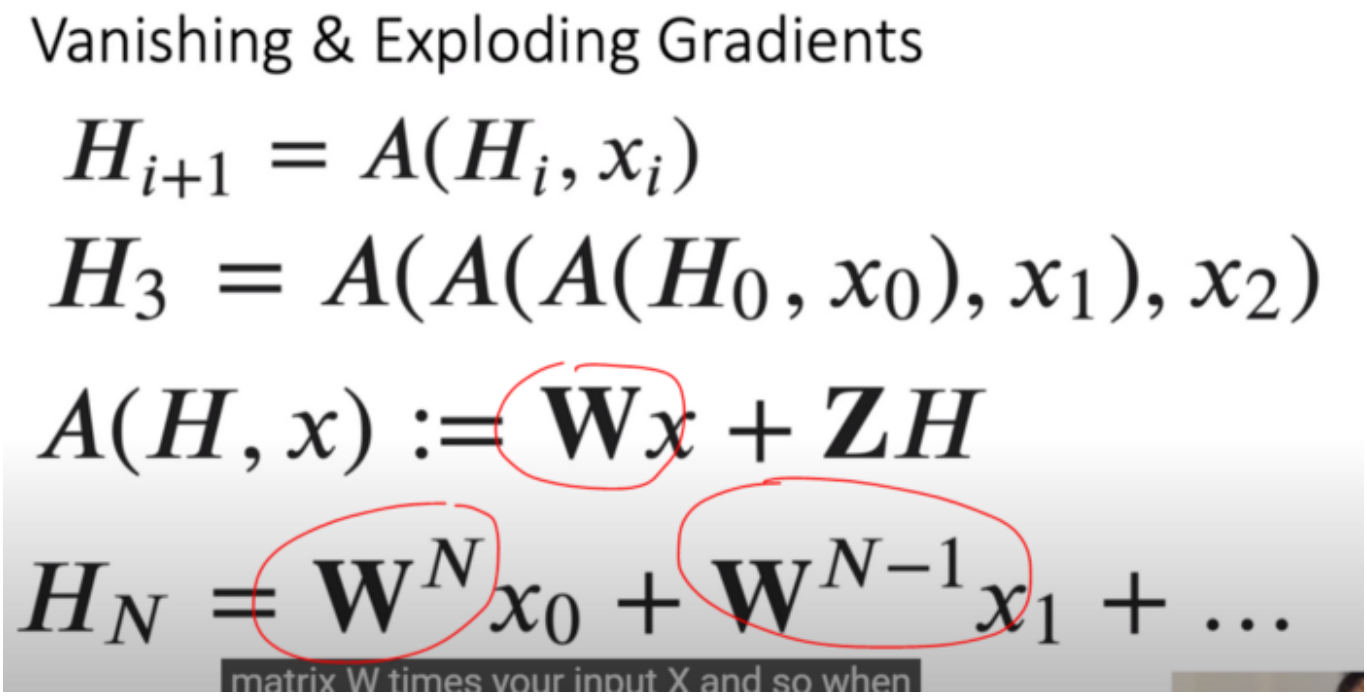


The traditional way to do this is to use a for loop in math. This is a good way, but it is not the optimal way.



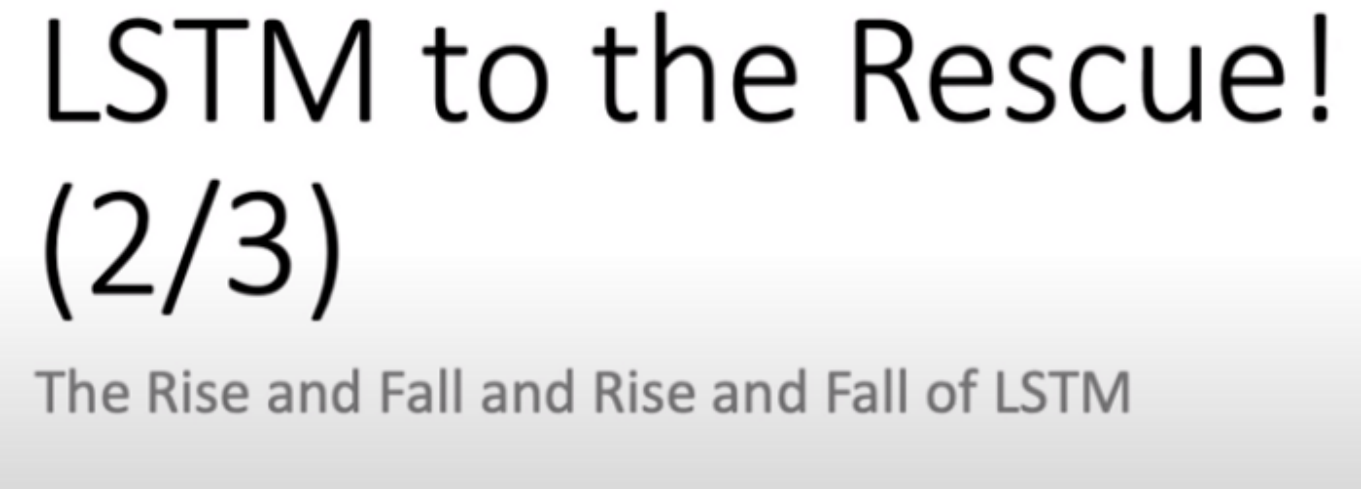
Big Data Jobs

Vanilla RNN the problem is vanishing as well as exploding gradient.

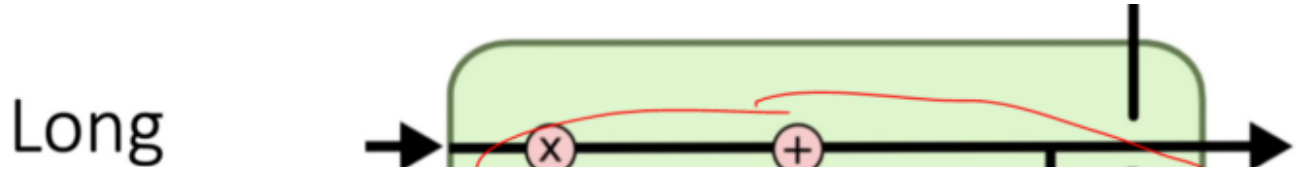


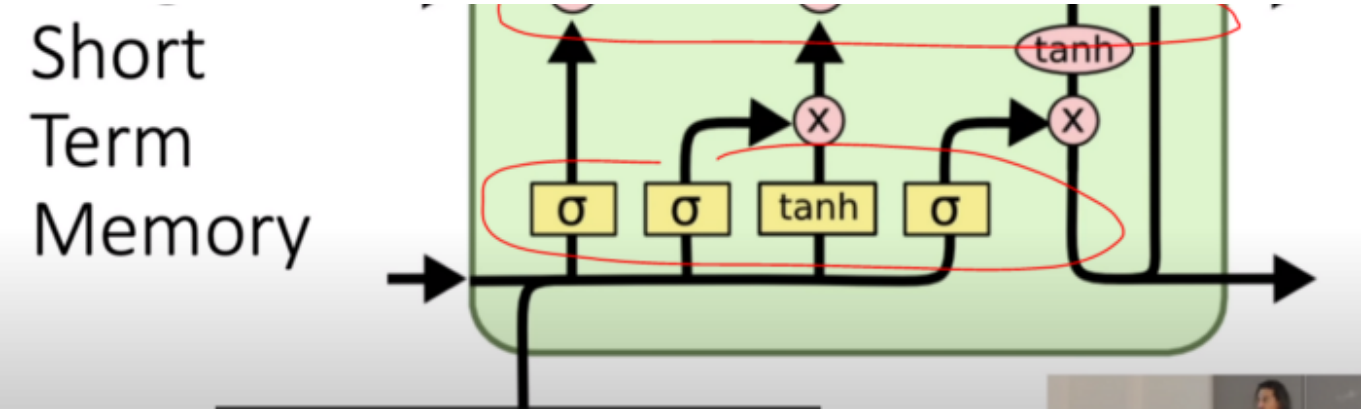
The gradient can vanish and this is not stable for training, the concept is good but in general, it is not practical.

The eigenvalue is what, makes this either good or not.

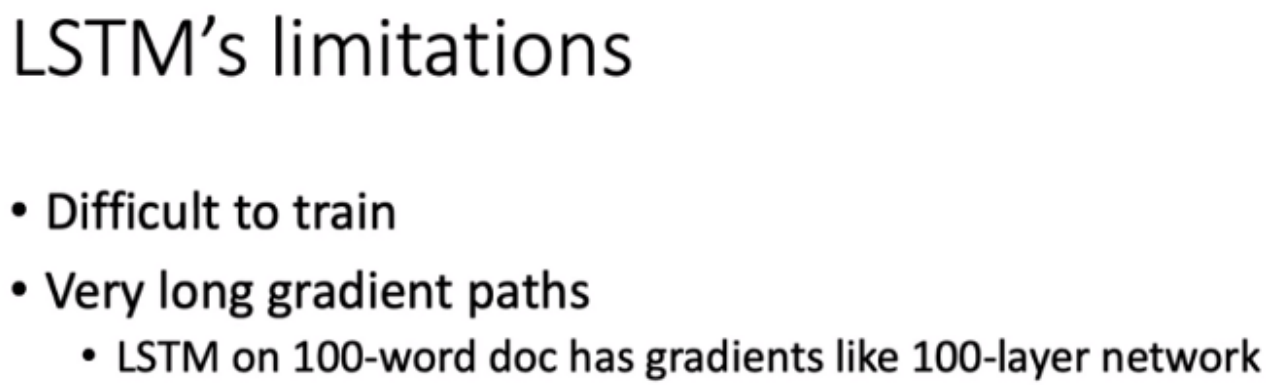


LSTM was the solution, it was a good solution. After the AI winter, it came back, a more advanced version of RNN. But now it seems like it is going away.

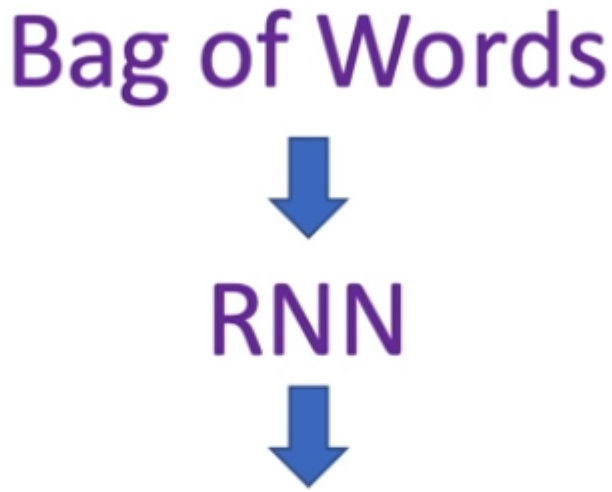




The math is pretty complex but very good! And there are many different kinds of operations and gates. It is like an RNN but with better control. Yet, difficult to train.



Transfer learning never really worked on this model, for language models. This is very bad since we are able to build upon other people's work. Such as BERT models.



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

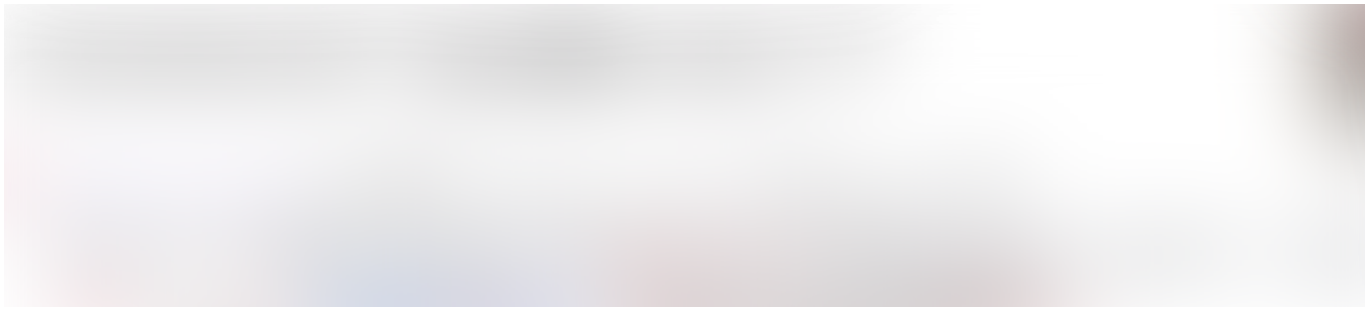
Top 4 Most Popular Ai Articles:

1. AI for CFD: Intro (part 1)
2. Using Artificial Intelligence to detect COVID-19
3. Real vs Fake Tweet Detection using a BERT Transformer Model in few lines of code
4. Machine Learning System Design

Still, quite a bit is going on, but the attention is the key part, and the way they do this is via all-to-all comparison. This is a great way.

We can actually see, what is going on in the model itself since we are able to visualize the attention. We need to translate THE, the next we need to look at the other part, one by one. And this is actually a good idea, the order is reversed.

Every output position we generate a query and from there we are going to get a relevancy score. Interesting. And this is all differentiable.



Still, there is a for operation, a list of tensors, one per token, and we are computing the value. Q, K, and V have all learned matrix.

A complex way to do this didn't know how much complexity is there for the NLP task. They are interactable and understandable.

And scaling this thing into a multi-head is not a hard thing to do. So it is scalable as well as transferable. But there is positional encoding, that brings this all together.

And they use different frequencies via sin/cos signals. This is how the system understands position. N^2 and this operation can be parallelized.

Additionally, we do not have to use S activation functions, since there is a dying gradient.

Those are some good reasons why RELU is really good, more robust, faster, and theoretically good too!

Some good points for deep learning. Both PyTorch and TensorFlow have an implementation.

And remember this is re-useable.

Depending on the use case LSTM still has its place.

Don’t forget to give us your 👏 !