dMila Talk

Hi everyone, my name is Giulia - I'm a research fellow at UCL. I've mostly worked using optimal transport in ML and understanding theoretical guarantees - so my main publications are also on the theory side. So with Joumana we agreed to present this paper today "Efficient sequence modelling …. " by …

The presentation is organized as follows:

1. **First I will state the motivation of their work and set up the model that they consider. Since the paper is heavily focused on the efficient computation of a quantity, I will make a small digression to explain where this quantity comes from and why we are interested in it.**
2. **Then I will proceed to discuss the details of the paper**
3. **And showcase some of their empirical results.**

Modeling and learning from sequential data is a fundamental problem in modern machine learning, underlying tasks such as language modeling, speech recognition, video processing, and reinforcement learning. A core aspect of modeling long-term and complex temporal dependencies is memory, or storing and incorporating information from previous time s

Just to set the stage, the problems that this models want to solve are long and implicitly continuous sequences .

Some examples are: medical time series like EEG, or speech waveform or energy forecasting. So we deal with data that is sampled from an underlying continuous process.
Speech data is sampled at 16000Hz, which means that for clip of just a few seconds, The sequences are tens of thousands of samples long.

Sequence model = sequence to sequence map composed of primitive layers where concretely each layer should operate on a sequence as input that is a sequence with a hidden dimension and outputs something with the same shape

RNN are models that summarize the information that they have seen up to a certain point in a hidden state - they are inherently sequential and suffer for vanishing gradient problems.
These and also CNN or continuous models struggle with LRD.

The model proposed is based on a state space model and at a high level this model has different representations, that correspond to different paradigms (continuous time, recurrence or convolution).

LRA : recent benchmark for long context sequence modelling. 6 tasks

SSM:
For each input you update latent or hidden state and the project to an output

Function approximation: the measure μ controls the importance of various parts of the input domain,

HiPPO specifies a class of certain matrices $A \in \mathbb{R}^{N \times N}$ that when incorporated into (1), allows the state x(t) to memorize the history of the input u(t).

Sketch of the algorithm that S4 introduce: it allows to reduce several polynomial factors in the runtime

Regarding the experiments:

Classification on images that have been flattened out and treated as sequence, where themodel does know the 2d structure. There are a few datasets here and a lot of prior work that they compare to. The main take home message is that while for sMNIST and pMNISt there is not much of a difference since we also don't include a standard deviation, for sCIFAR the improvement is significant, more than 15% higher accuracy than the other benchmarks. It's also interesting that the performance is not too far from the performance of a ResNet - where the 2D structure is encoded and exploited.

The S4 model tested on speech classification task: the task is called keyword spotting and consists in 1 second audio clips that have to be classified into words. Speech is hard because of the high frequency, 1 second clips correspond to sequences of length

16000.

Speech pipelines normally don’t use raw data but require some feature engineering

And one option is to use MFCC that reduces the size from 16000 to 160. With these features the results are reasonable with a range of different models and S4 positions in the higher part of the spectrum.

When using raw data however, it seems that the only model that gives good results is the WaveGan-D model that is specifically designed for this purpose and that the authors point out is 88X larger of S4. Also it outperformes the performance on the MFCC features.

The testing, without retraining was done on data sampled with 8K Hz. While Other model suffer, S4 performes well and the intution is that since it tries to model a continuous process it is not affected by how the sequences are sampled from such model.

Long RAnge arena:

Most interesting result is on path-X

The same dataset was used also for benchmarking

Finally, they provided results on text in particular in WikiTest-103 dataset. Here I think there is an extra step in the sense that they take a transformers backbone and they swap the attention layer with the S4. I didn’t go into the details of this but the main takeaway is that It is only a less that one point perplexity off even though it is attention free and it also provides an advantage in terms at generation time.

GRU:

LSTM:

Alex Graves LSTM paper
Speech Recognition with Deep Recurrent Neural Networks: http://arxiv.org/abs/1303.5778

Alex Graves (43 pages) paper: http://arxiv.org/abs/1308.0850

Colah's blog: Understanding LSTM Networks -- colah's blog

Transformer

1. Hey, My name is Akshay Shelke I was excited to learn the transformers and BERT due to its wide applications and problem use cases I deep-dived into it the resources that were helpful to me,I would like to share a few
   1.https://jalammar.github.io/illustrated-transformer/
   -Jay has Easily explained the Transformers also read his BERT blog
   2.https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
   Try Searching similar blogs on Medium and Kaggle

2. For basics visit the CampusX YouTube channel it's in the Hindi Language for English visit Krish Naik and codebasics…(I recommend watching only BERT and Transformers videos for basic understanding for deep knowledge visit the Standford online channel)
https://www.youtube.com/@campusx-official
https://www.youtube.com/@codebasics
https://www.youtube.com/@krishnaik06

3. Go through Kaggle and Analytics Vidya blogs.

BERT

BERT Fine-Tuning Tutorial with PyTorch: https://mccormickml.com/2019/07/22/BERT-fine-tuning/

GPT

lş