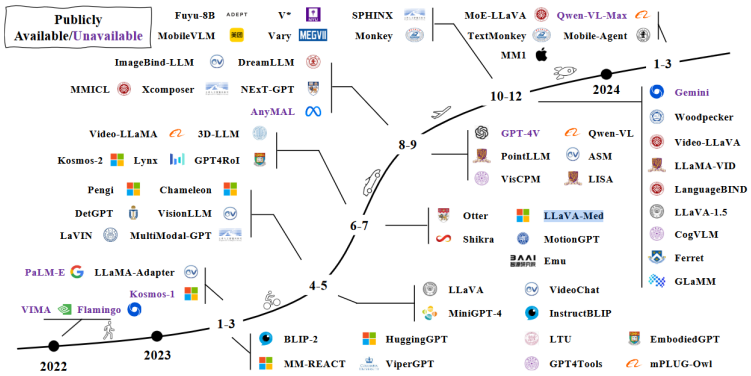
Multimodal LLMs for Time Series

Michael Staniek

Department of Computational Linguistics
Heidelberg University

October 14, 2024

Multimodality is Hype



The Data

https://github.com/harsh19/TRUCE/tree/main[2]

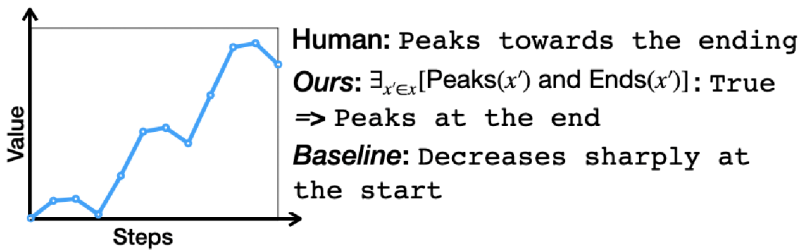


Figure: The dataset that I want you to use

The general Idea

- The repository contains 2 different datasets. One is completely synthetic, the other one one is works on Stocks
- "annotations": [["Flattens off towards end."], ["Rises sharply in the middle."], ["Increases in the middle"]], "series": [4, 6, 6, 3, 3, 4, 17, 30, 30, 31, 30, 30]
- I want you to try out the best Encodings for both dataset
- Does it make more sense to:
  1. Use a encoder-decoder model?
  2. Use a LLM and give the numbers directly as input?
  3. Some other multimodal architecture?

Other architecture

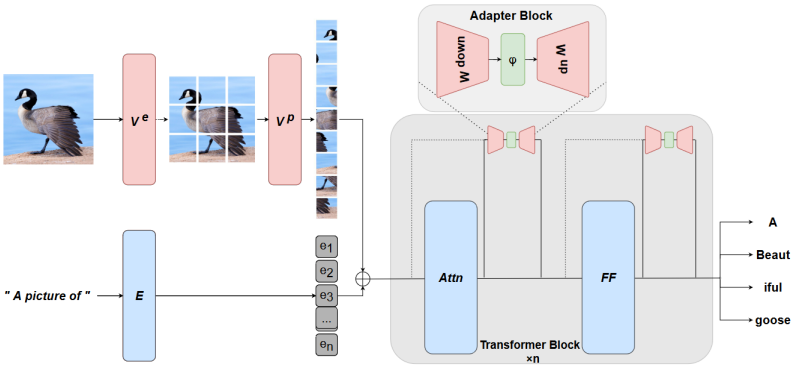


Figure: Input the time series as an image or inputting the time series as a time series with a time series encoder, which is better?[1]

Deliverables

- Implement Encoder-Decoder and train it from scratch on this dataset
- Finetune a LLM on this dataset
- Implement the MAGMA architecture and finetune it with a LLM of your choice (doesn't have to be the largest LLaMA model)
  - Represent data as image directly and use an image encoder.
  - Use a seperate time series encoder.
- Evaluate the performance of all models. Use appropiate measurements (BLEU, CHRF...).

The End

References

[1] Constantin Eichenberg et al. “MAGMA - Multimodal Augmentation of Generative Models through Adapter-based Finetuning”. In: *CoRR* abs/2112.05253 (2021). arXiv: 2112.05253. URL: https://arxiv.org/abs/2112.05253.

[2] Harsh Jhamtani and Taylor Berg-Kirkpatrick. “Truth-Conditional Captioning of Time Series Data”. In: *EMNLP*. 2021.