New Sources of Inspiration

Over the past several days, I have been studying the transformer architecture and building small projects that demonstrate its efficiency in generating novel data. Among the many resources I explored, the most valuable was the groundbreaking paper *“Attention Is All You Need”*, published by a team of researchers at Google. This innovation laid the foundation for the recent explosion in deep learning, and has been especially pivotal in training large language models like ChatGPT.

Furthermore, I found Andrej Karpathy’s excellent YouTube series, created by the former OpenAI co-founder extremely helpful for gaining hands-on coding experience with the structure and subcomponents of the decoder transformer model. The educational series by 3Blue1Brown was critical, offering amazing visualizations of complex deep learning concepts that greatly enhanced my understanding.

I have learned a large amount in the past few days, but this is only the beginning of my deep learning journey.

What I Have Learned

The most significant innovation in the transformer architecture is the self-attention mechanism. Unlike previous models I have trained, self-attention allows the tokens in the training data to communicate with one another, creating a much more effective learning process that feels almost magical.

The mechanism operates through the interaction of three matrices: query, key, and value. The query and key matrices interact in a way that lets tokens draw information from previous tokens, effectively allowing the model to "ask questions" about its own structure. The responses to these questions are captured through the key matrix, and the relationships are computed using their dot product. The value matrix then reinforces these results, enabling tokens to shift meaningfully from their original embeddings.

I find the relationship between these three matrices astonishing, and inspirational to the level of research I hope one day to produce

What's Next?

After developing a better understanding of the transformer, I hope to explore more frontier research on the different variations of the model. When I am comfortable, I will begin my exploration in the world of generative AI for the purpose of AMP production, combining my passion for bioinformatics and deep learning. In the mean time, I will continue to build small projects, poking and prodding the limits of the transformer architecture, and my poor Apple M2 processor.

Notebook snippet from "Character-Level_LM_Transformer", inspired heavily by Andrej Karpathy:

```python
# Sampling generation trained on the works of Shakespeare:
idx = torch.zeros([1, 1], dtype=torch.long)
print(decode(m.generate(idx, 1000)[0].tolist()))
```

```
HERMIONE:
Good lie.

KE VINGHAM:

QUEEN MARGARET:
Here is forbeaid; let's respite what correction?

GLOUCESTER:
Mine hold of this, patient his royal livew.

QUEEN MARGARET:
Being noble lord field; Petreach,
For them beg so; as the people's officer
Supplaint these princivation
Not Richshor,
And for the toldmast Apollo's bellock,
As answer seved as three but our Chrereford's
lieferant pardons thousand mindstress'd kings.
Now Traitors, Bucut, while virtuous he two
Is every counsel for ut hinder at With had,
Betward to heir yeld wear at Tuney.
Go, be I will gone with our burain?

RIVERS:
Now grantal triumphans: by your fanty his eambans:
I am tell there whenty man, for that hangs,
Whill to talk us will without mine still remembers;
Bere those needs it fears:
As miseasure you were gone a farmers to
false: Ifter he will in the world poof; Which
to London keep itself
and other offinnen the ploters.
```