

Transformers

Notes on various aspects of Transformers

Contents

1	The Architecture	1
2	Memory	1
A	Conventions and Notation	1
B	To-Do	1

1 The Architecture

The Transformers architecture [1], which dominates Natural Language Processing (NLP) as of July 2023, is a relatively simple architecture. There are various flavors and variants of Transformers, but we start by describing the decoder-only versions which underly the GPT-X models [2–4].

2 Memory

```
def hello_world(args):  
    print("hello world this is a test")
```

A Conventions and Notation

We loosely follow the conventions of [5] and denote the main Transformers parameters by:

B To-Do

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” [arXiv:1706.03762 \[cs.CL\]](#). **1**
- [2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog* **1** (2019) no. 8, 9. **1**
- [3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” [arXiv:2005.14165 \[cs.CL\]](#).

- [4] OpenAI, “Gpt-4 technical report,” [arXiv:2303.08774](#) [cs.CL]. 1
- [5] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, and B. Catanzaro, “Reducing activation recomputation in large transformer models,” [arXiv:2205.05198](#) [cs.LG]. 1