Transformer from Scratch

Marcel Fedosejew, Eric Loßkamp, Jasper Orthen

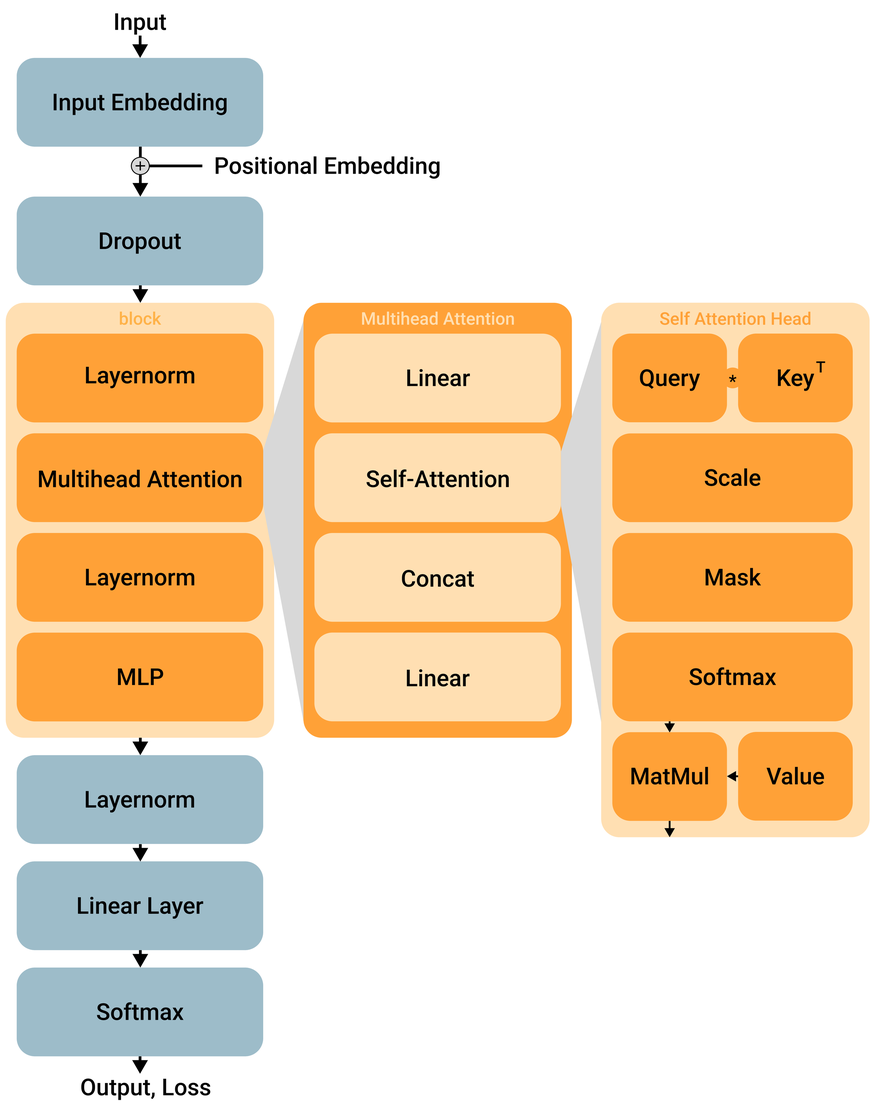
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Introduction

Transformers are often treated as black-box architectures.We implemented a decoder-only transformer from scratch (including all backward passes) to investigate:

- How training dynamics behave without high-level frameworks
- The impact of dropout on convergence
- Whether decreasing loss reflects improved generation quality

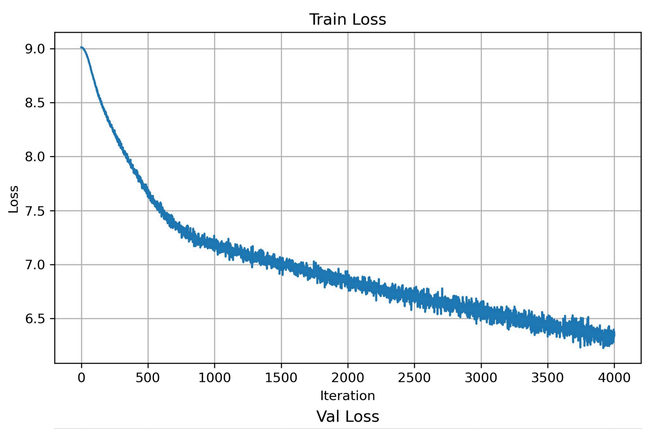
Architecture

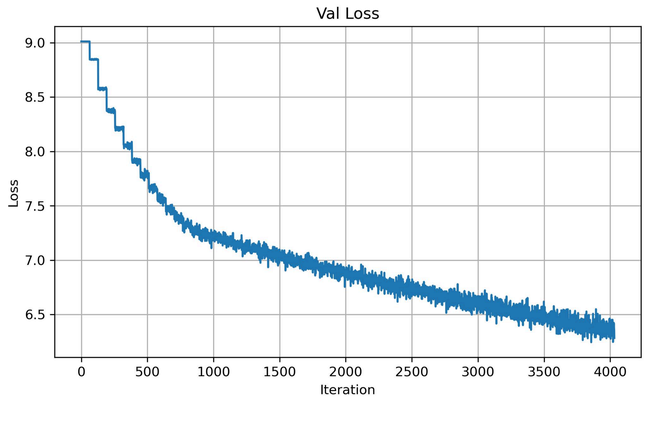


Implementation Challenges:

- Gettings the right shapes of inputs and outputs
- finding the issues of wrong calculations
- running on the provided hardware

Training



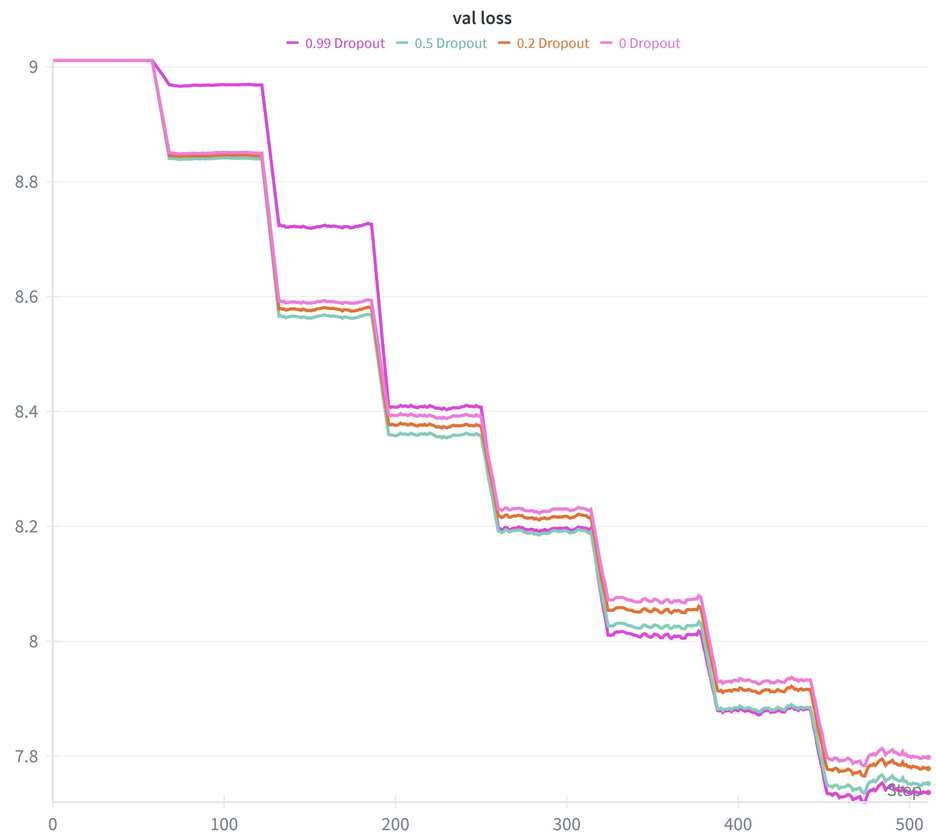


Settings
- Tokenizer V2
- batch size: 16
- Context Length: 256
- dropout: 20%
- learning rate: 1e-3
- params: 10,6m

Training Setup:
- 2x RTX 2080Ti
- 100 minutes for 500 epochs
- 145mb weights

Observations:
- 64 iter “steps”
- visible convergence

Dropout Study



Effect of Dropout on Convergence:

- High dropout doesn’t stop convergence
- The impact of dropout on the loss is minimal
- Until the 200th iteration, 99% dropout performs worse
- at the end 99% dropout has the lowest loss

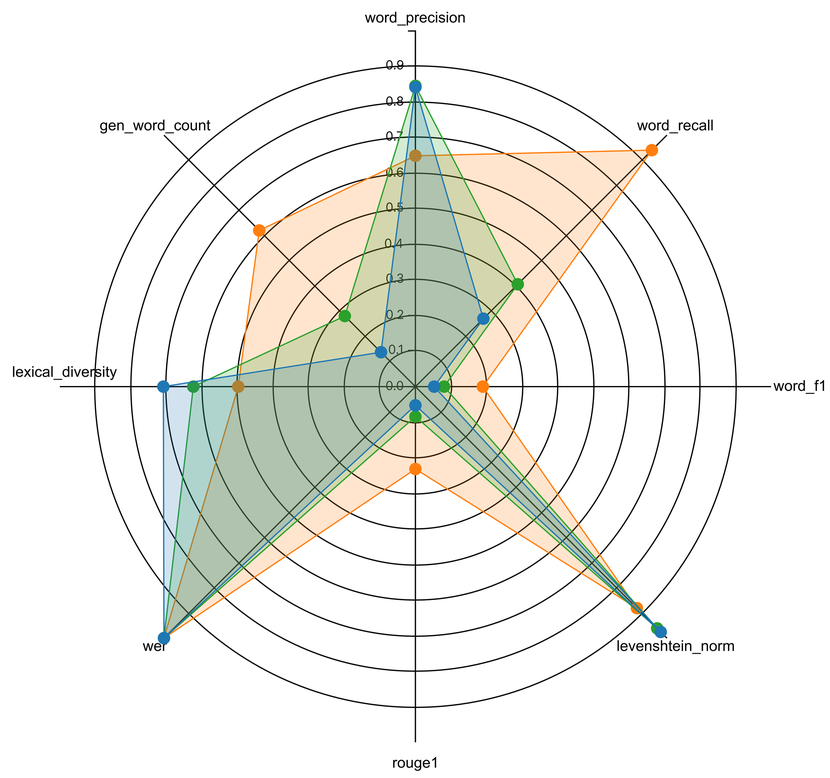
Generation Quality

Observations:

- Word count shrinks with higher iteration count
- word diversity and precision improve
- word error rate stays high

Setup:

- 65, 7104, 13952
- 500 tokens
- temperature 0.8
- init-sequence



65. Epoch

's-- uns zu nur. dem dieSo zuDer, und
den der uns den derDer uns dich [!--
ihr: er dich?--Was, von's--DieDer ein
zu - uns zu?-- zum sie sie [ihr ein
der? so nur,Die? zumDie sie der
die!-- einenWas'sDie sie einen einen
- zu zuWas? uns die derDer
- er dich sie
den von den's und [sie einen, nun--
von uns [ein die - nicht [
zuDer die eine den der nun
du mich eine zu, derDie den und
uns das? und -Die - nun das! dich;
wir
! dich
's und,DieDer sie zu: vonDer nur?--
zu - dich einen.] eine?Die dich]. von -
So der und ...

7104. Epoch

Und dir der .
Wenn und mit,

Und mir die

Das ist,
Die und was ich nicht..
Das das
[
So dich und er sich,
Die du,

Es du,

FAUST,, ...

13952. Epoch

Ich mich zu sein, so
Der Zeit,

Sie wird's zu

Wie er dem
Die Rau! du den Herz?
Es,

Ich war?
Doch das Mann.
FAUST.

Mit dem Täs!
...

Conclusion & Outlook

Conclusion
- A transformer can be reliably implemented from scratch with careful gradient handling
- Dropout minimaly affects convergence and generalization
- Training loss alone is insufficient to evaluate generation quality

Outlook: Larger models, alternative/larger datasets, and improved evaluation metrics.