



# Char-level decoder-only transformer on Shakespeare data

Project of Deep Learning course

**Marco Carotta**

16 December 2024



# Project (and repository) structure

Outline



# Project (and repository) structure

## Outline

- `dataset.py`



# Project (and repository) structure

## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection



# Project (and repository) structure

## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection
- `model.py`



# Project (and repository) structure

## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection
- `model.py`
  - Multi head (self-)attention transformer from "Attention is all you need", with some differences



# Project (and repository) structure

## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection
- `model.py`
  - Multi head (self-)attention transformer from "Attention is all you need", with some differences
- `train.py`



# Project (and repository) structure

## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection
- `model.py`
  - Multi head (self-)attention transformer from "Attention is all you need", with some differences
- `train.py`
  - Adam, schedulers, single GPU (Kaggle), metrics tracking





# Project (and repository) structure

## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection
- `model.py`
  - Multi head (self-)attention transformer from "Attention is all you need", with some differences
- `train.py`
  - Adam, schedulers, single GPU (Kaggle), metrics tracking
- `generate.py`



# Project (and repository) structure

## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection
- `model.py`
  - Multi head (self-)attention transformer from "Attention is all you need", with some differences
- `train.py`
  - Adam, schedulers, single GPU (Kaggle), metrics tracking
- `generate.py`
  - Temperature and top p-nucleus sampling



# Project (and repository) structure

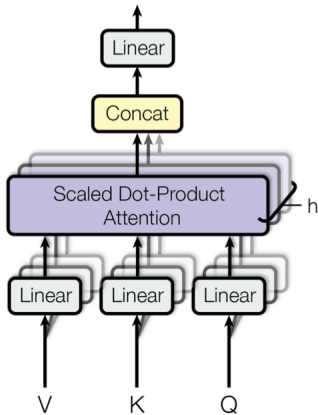
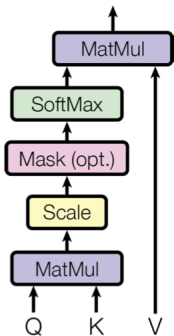
## Outline

- `dataset.py`
  - Char-level tokenizer, preprocessing (lowercase, remove char with low frequency, punctuation)
  - Noise injection
- `model.py`
  - Multi head (self-)attention transformer from "Attention is all you need", with some differences
- `train.py`
  - Adam, schedulers, single GPU (Kaggle), metrics tracking
- `generate.py`
  - Temperature and top p-nucleus sampling
- some scripts and folders with models or figures



# Multi-head attention mechanism

## Outline



From "Attention is all you need"



# Transformer Block

## Outline

Difference:

- Decoder-only



# Transformer Block

## Outline

Difference:

- Decoder-only
- Pre-layer normalization, following the idea in "On Layer Normalization in the Transformer Architecture"

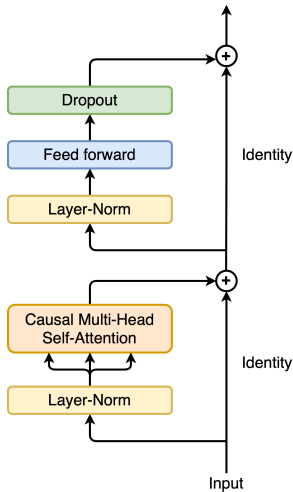


# Transformer Block

## Outline

### Difference:

- Decoder-only
- Pre-layer normalization, following the idea in "On Layer Normalization in the Transformer Architecture"

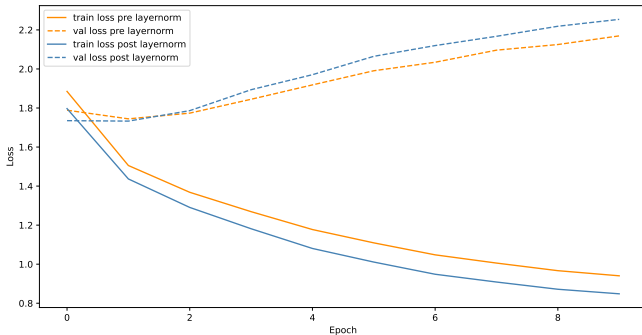




## Training Outline

### Hyperparameter:

- batch size = 256
- context length = 128
- embed dim = 512
- num heads = 8
- num blocks = 8
- ff hidden dim = 1024
- total par = 17 millions
- lr = StepLR
- time = 8h







## Generation

Outline

Temperature: 0.2

clown:

indeed, if it be too much blood, your  
pratest, as it were, farewell.

archidamus:

i think so. kill'd!

she i kill'd! i did so: but thou strikest me  
sorely, to say i did; it is a power to die.

nurse:

Temperature: 1.0

bear thy mother's blood this day should  
choose  
the citizens of choice; she cannot live,  
i doubt but rinnocency be will still.

cominius:

you have made good work!

menenius:

what ever this?

cominius:

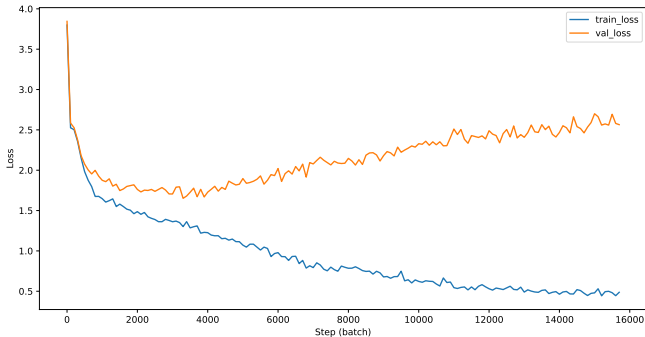
yes, marcius.



## Training Outline

### Hyperparameter:

- lowercase
- context length = 128
- embed dim = 512
- num heads = 8
- num blocks = 10
- ff hidden dim = 1536
- total par = 25 millions
- lr =  $3e-4$  (no sched.)
- time = 2h





## Generation

### Outline

Temperature: 0.2

lady capulet:

well, think of marriage now;  
younger than you,  
here in verona, ladies of  
esteem,  
are made already mothers:  
by my count,  
i was your mother much  
upon these years  
that you are now a maid.  
thus then in brief:

Temperature: 0.6

second citizen:

i pray you, what he hath  
done famously, he did  
it to that end: though  
soft-conscienced men can  
be  
content to say it was for his  
country he did it to  
please his mother and to be  
partly proud; which he

Temperature: 1.0

you shall not then be spoke  
to that must be cool'd  
for bolingbroke says that  
ffor t.ee.

king richard ii:

doubly divorced! bad men,  
you violate!  
a thriftuon traitor to  
marcius poor in, that  
when the sequel had not  
poor,



# Conclusions

## Outline

- Small models but meaningful text



# Conclusions

## Outline

- Small models but meaningful text
- Overfitting
  - Limited computational resources



# Conclusions

## Outline

- Small models but meaningful text
- Overfitting
  - Limited computational resources



# Conclusions

## Outline

- Small models but meaningful text
- Overfitting
  - Limited computational resources

Thanks for the (self-)attention



# Conclusions

## Outline

- Small models but meaningful text
- Overfitting
  - Limited computational resources

Thanks for the (self-)attention

## Bibliography

- Ashish Vaswani et al. (2023). Attention Is All You Need. [link]
- Ruibin Xiong et al.(2020). On Layer Normalization in the Transformer Architecture. [link]
- Andrej Karpathy, NanoGPT GitHub repository [link]