Эксперимент 1 (начальный):

Слова были токенизированы с помощью *токенизатора* BPE:

- vocab_size=4000

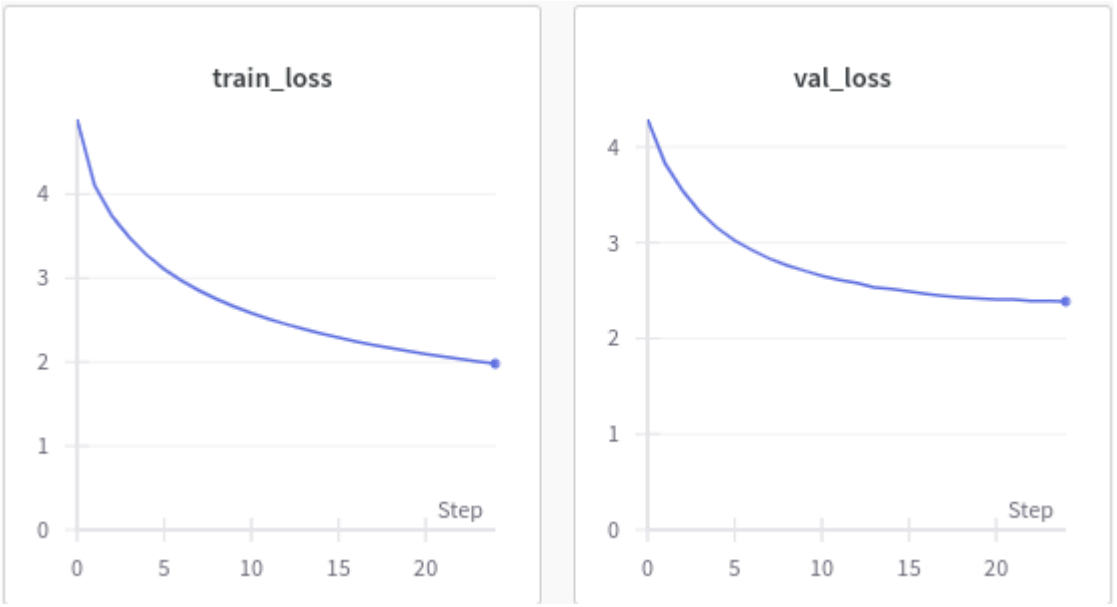
В качестве *архитектуры* был взят nn.Transformer с параметрами:

- d_model=512,
- nhead=8,
- num_encoder_layers=5,
- num_decoder_layers=3,
- dim_feedforward=512.

Во всех следующих экспериментах эти параметры не менялись.

Использовался *оптимизатор* Adam

- lr=1e-4,
- betas=(0.9, 0.98),
- eps=1e-9.



Ссылка на wandb: https://wandb.ai/ntyazh-team/padding/runs/3oa4qaf4?nw=nwuserntyazh

Видно, что после 15 эпохи модель начинала переобучаться, так что было решено добавить дропаут в трансформер.

Эксперимент 2

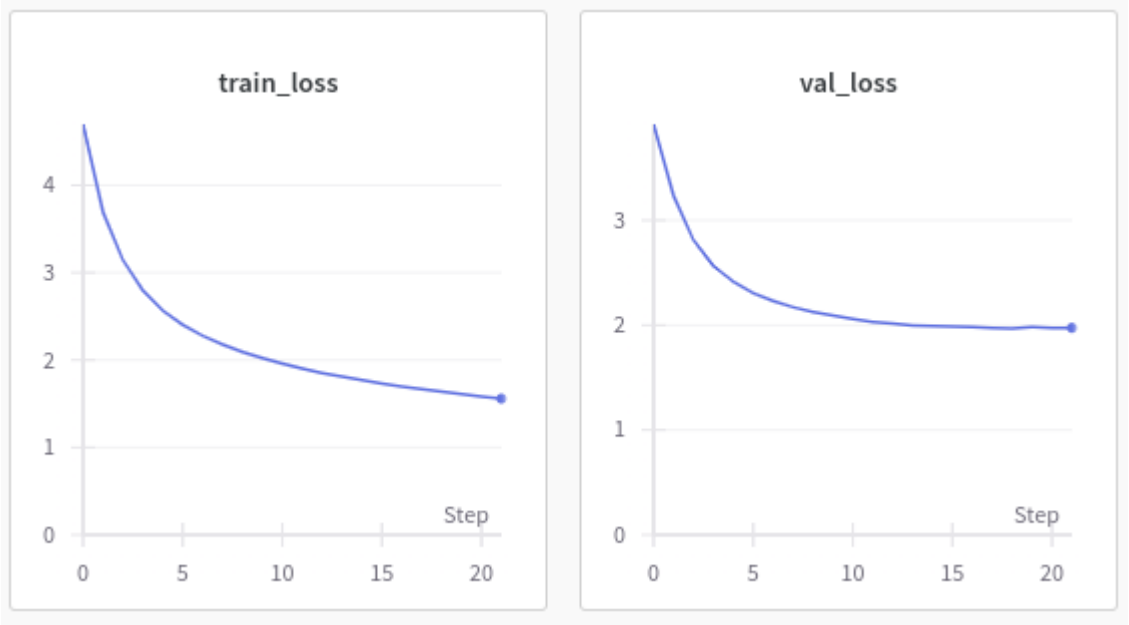
Архитектура nn.Transformer

- dropout=0.1.

Оптимизатор Adam.

Как видно, это несильно спасло от переобучения, однако лосс на валидации уже стал ниже 2. Далее было решено попробовать добавить lr-scheduler CosineAnnealingWarmRestarts, потому что для трансформеров обычно используют

шедулеры с warm restarts, а cosine annealing – один из самых популярных шедулеров в принципе.



Ссылка на wandb: https://wandb.ai/ntyazh-team/dropout/runs/iw4vmhaf?nw=nwuserntyazh

Эксперимент 3

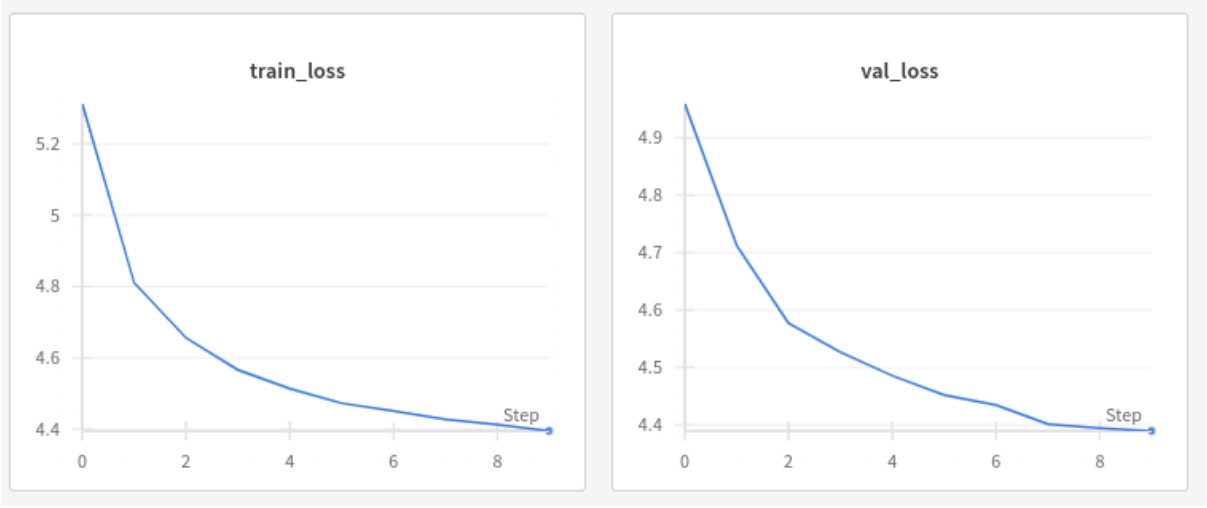
Архитектура: Transformer

Оптимизатор: Adam

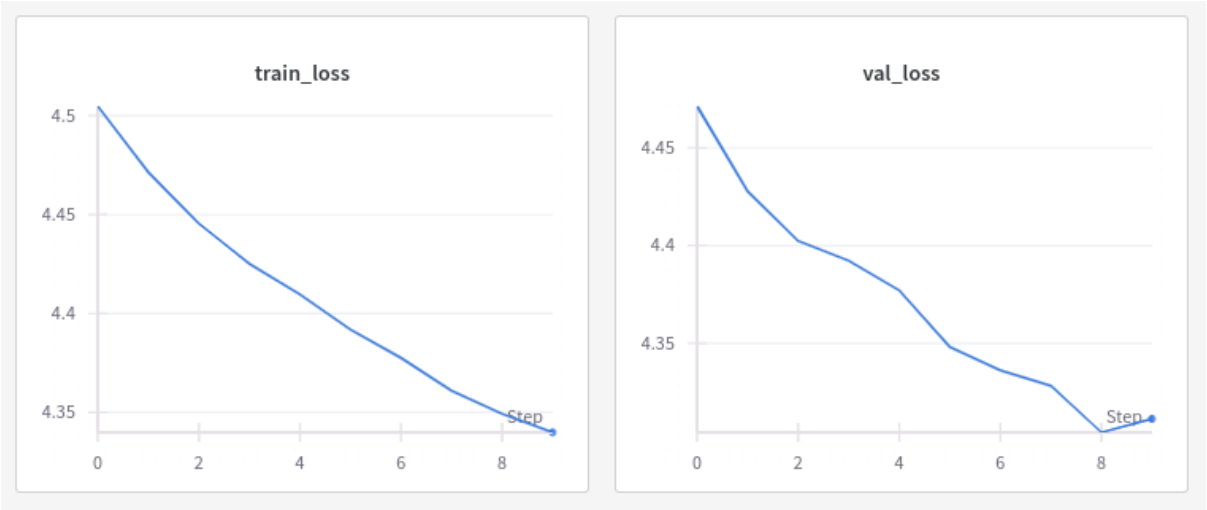
LR-scheduler: CosineAnnealingWarmRestarts

- T_0=5,
- T_mult=1.

Первые 10 эпох:



Вторые 10 эпох:



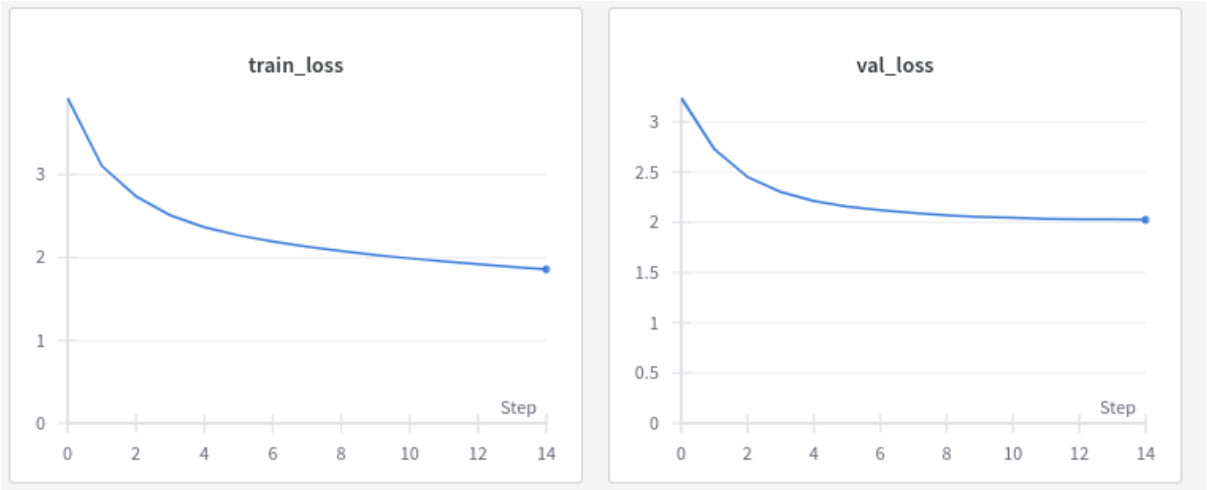
Как видно, модели просто стало гораздо сложнее учиться, поэтому было принято решение не брать шедулер вообще.

Эксперимент 4

Далее я решила попробовать пословный *токенизатор* word:

- vocab_size=4000

Без шедулера и с предыдущими параметрами трансформера и оптимизатора.



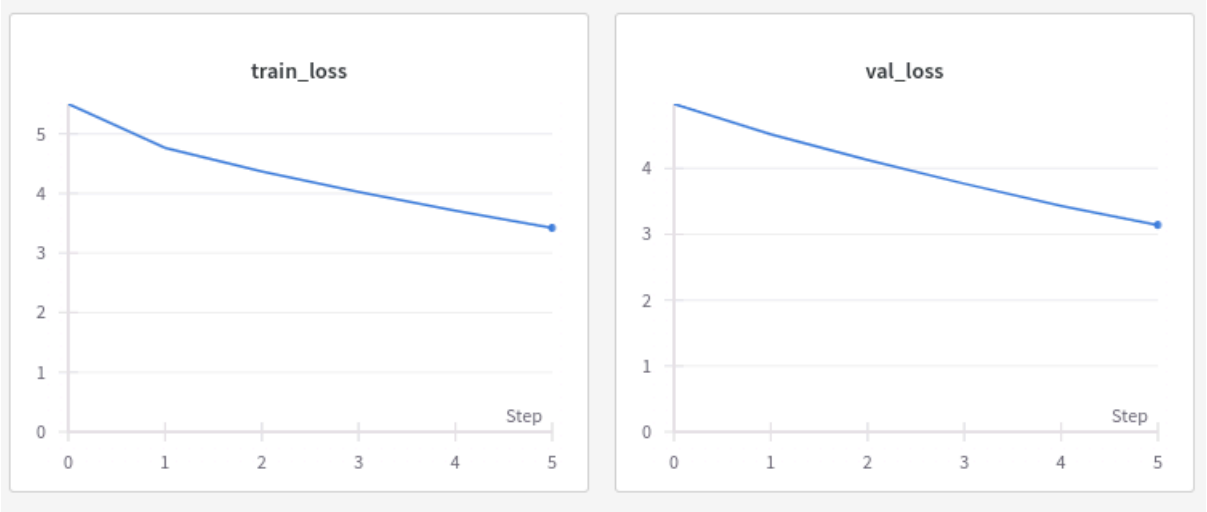
Ссылка на wandb:
https://wandb.ai/ntyazh-team/word_tokenizer/runs/2ve4ovpy?nw=nwuserntyazh

Результат получился хуже, чем с токенизатором BPE (эксперимент 2), так что было принято решение оставить BPE токенизатор.

Эксперимент 5

В одной статье про трансформеры (https://arxiv.org/pdf/1908.11365.pdf) я прочитала, что обычно их параметры инициализируют равномерным распределением Xavier uniform distribution.

Так что я добавила эту инициализацию к лучшей из имеющихся моделей (т.е. из эксперимента 2)



Ссылка на wandb:
https://wandb.ai/ntyazh-team/uniform_init_bpe/runs/mz4f17xq?nw=nwuserntyazh

Однако, вопреки ожиданиям, эта начальная инициализация не то что не сделала начальные лоссы меньше – она их ухудшила, так что я решила не брать её в итоговую версию.

Итоговая версия (эксперимент 2)

Токенизатор: BPE

- vocab_size=4000

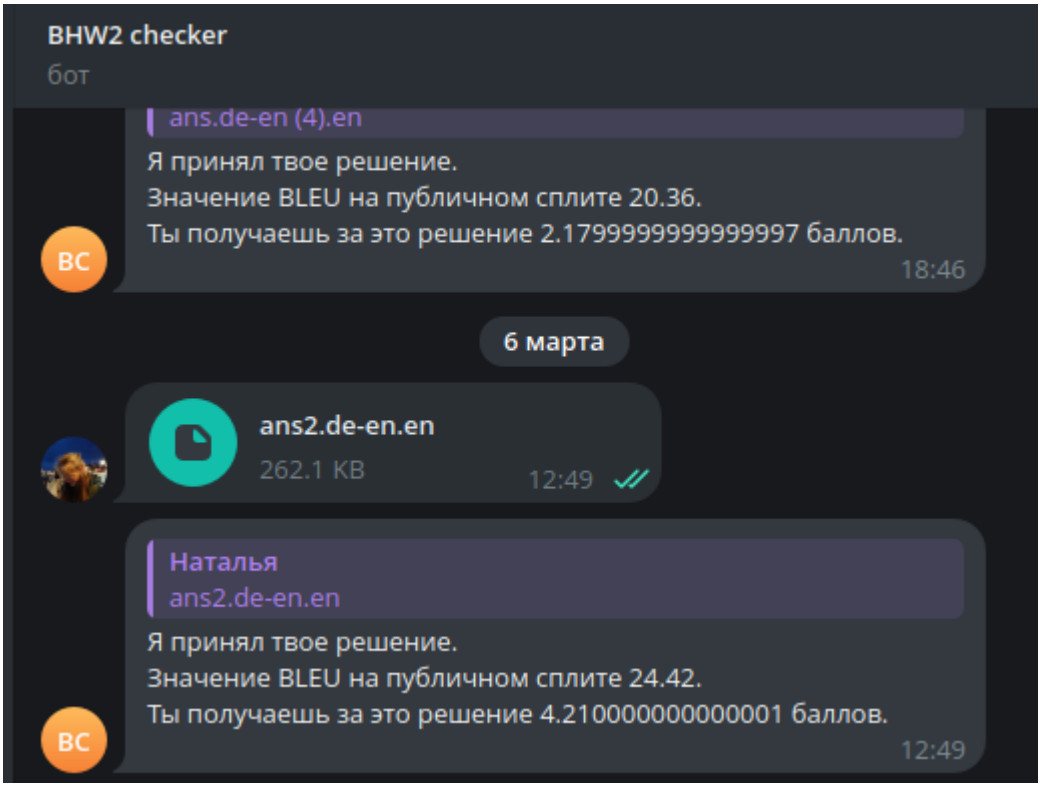
Архитектура: nn.Transformer

- d_model=512,
- nhead=8,
- num_encoder_layers=5,
- num_decoder_layers=3,
- dim_feedforward=512,
- dropout=0.1.

О*птимизатор* Adam

- lr=1e-4,
- betas=(0.9, 0.98),
- eps=1e-9.

BLEU (на публичном сплите) получился равным 24.42:


BHW2 checker
бот
ans.de-en (4).en
Я принял твое решение.
Значение BLEU на публичном сплите 20.36.
Ты получаешь за это решение 2.1799999999999997 баллов.
18:46
6 марта
ans2.de-en.en
262.1 KB
12:49
Наталья
ans2.de-en.en
Я принял твое решение.
Значение BLEU на публичном сплите 24.42.
Ты получаешь за это решение 4.210000000000001 баллов.
12:49