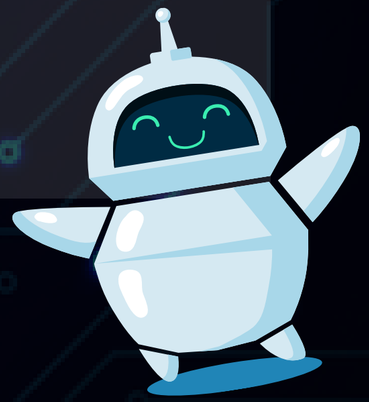
WHISPER

Boosting the model



Makarov Igor
Sokolov Pavel
Alshevskiy Dmitriy

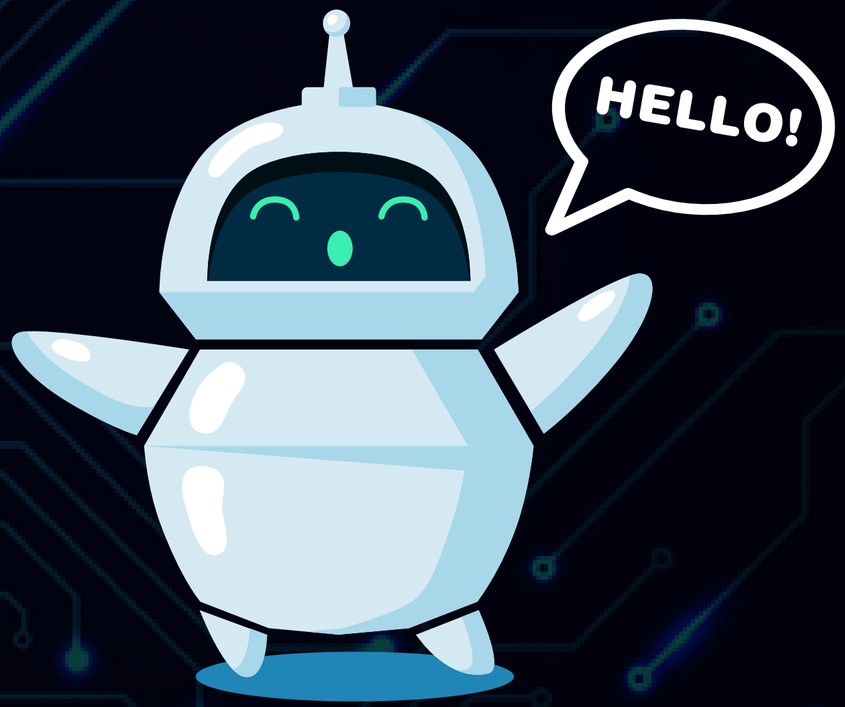
INTRODUCTION

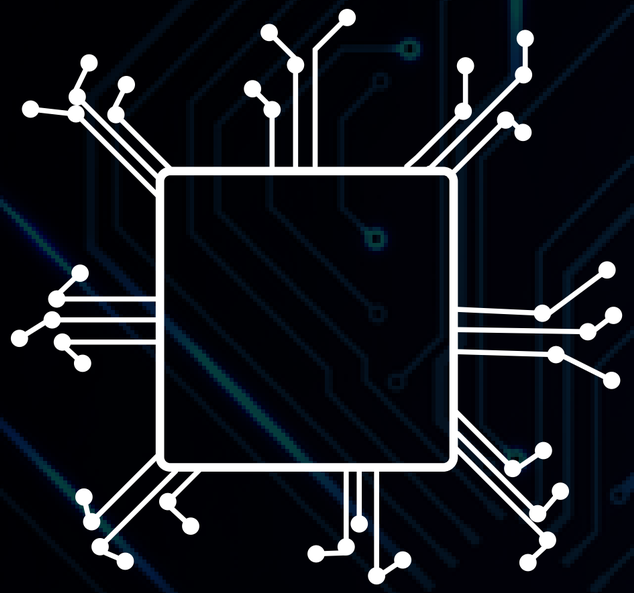
```
WhisperForConditionalGeneration(
  (model): WhisperModel(
    (encoder): WhisperEncoder(
     (conv1): Conv1d(128, 1280, kernel_size=(3,), stride=(1,), padding=(1,))
       (conv2): Conv1d(1280, 1280, kernel_size=(3,), stride=(2,), padding=(1,))
            (embed_positions): Embedding(1500, 1280)
            (layers): ModuleList(
            (0-31): 32 x WhisperEncoderLayer(
                (self_attn): WhisperAttention(
                    (k_proj): Linear(in_features=1280, out_features=1280, bias=False)
                    (v_proj): Linear(in_features=1280, out_features=1280, bias=True)
                    (q_proj): Linear(in_features=1280, out_features=1280, bias=True)
                    (out_proj): Linear(in_features=1280, out_features=1280, bias=True)
                )
                (self_attn_layer_norm): LayerNorm((1280,), eps=1e-05, elementwise_affine=True)
                (activation_fn): GELUActivation()
```

BASELINE

Sberdevices-golos-10

Device	WER	CER	Avg Time per Audio (s)
CPU	0.4403	0.1583	8.7137
GPU	0.4403	0.1583	0.8386





PROFILING

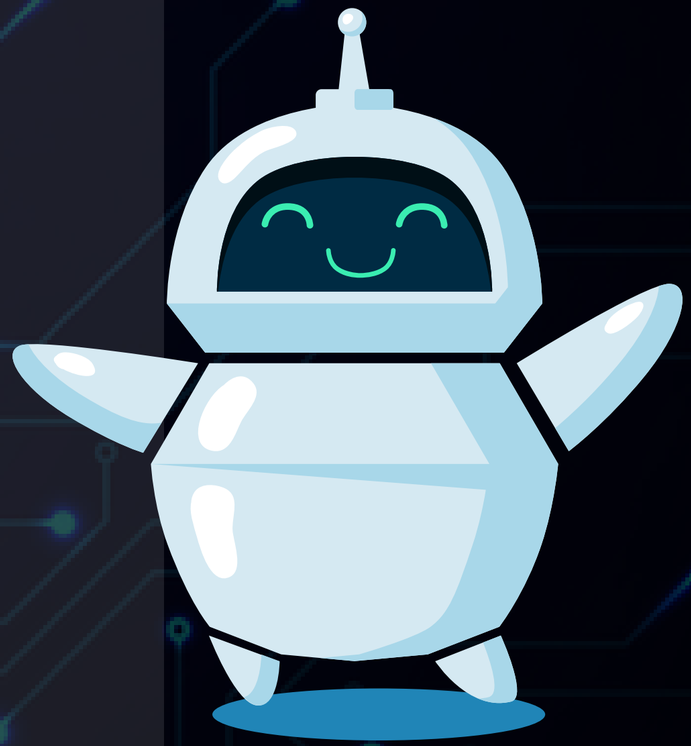
CPU:
aten::addmm типа output = beta * input + alpha * (mat1 @ mat2)

GPU
aten::_efficient_attention_forward
fmha_cutlassF_f32_aligned_64x64_rf_sm80

PRUNING

- Магнитудный прунинг. Отработал на CPU и GPU, значительного ускорения добились только на видеокарте (0.8 -> 0.6 секунд на запись)
- L2 прунинг. Разрушил модель при проходе, решено не использовать.
- Варианты прунинга с удалением весов реализовать не удалось.

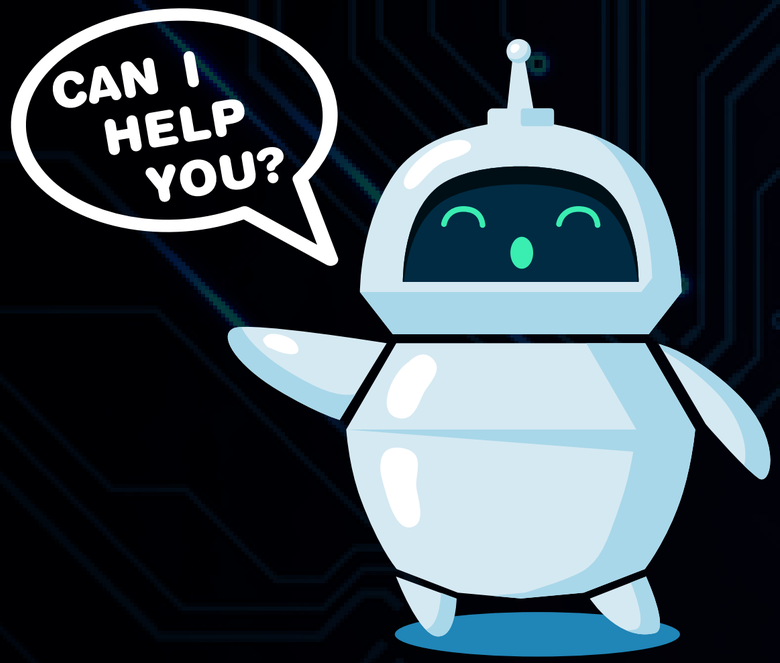
Pruning rate	WER	CER	Avg Time per Audio (s)
GPU base	0.4403	0.1583	0.8386
GPU 0.81	0.4410	0.1619	0.8386
CPU base	0.4403	0.1583	8.7137
CPU 0.81	0.4410	0.1619	7.4373



QUANTIZATION BASELINE

- torch.compile
- Static PTQ
- Dynamic PTQ
- BitsAndBytes
- torch.autocast

Quant	WER	CER	Avg Time per Audio (s)
GPU base	0.4403	0.1583	0.8386
GPU q	0.4472	0.1586	3.0069
CPU base	0.4403	0.1583	8.7137
CPU q	0.4740	0.1664	5.8057



DYNAMIC QUANTIZATION CONFIGS

Quant	WER	CER	Avg Time per Audio (s)
CPU base	0.4403	0.1583	5.4567
Int8AInt8W	0.4480	0.1582	4.8465
Int8AInt4W	0.4498	0.1610	10.5658

FINAL RESULTS

Configuration	WER	CER	Avg Time per Audio (s)
GPU base	0.4403	0.1583	0.8386
CPU base	0.4403	0.1583	8.7137
CPU q	0.4740	0.1664	5.8057
CPU p 0.81	0.4410	0.1619	7.4373
CPU q+p+c	0.4815	0.1664	5.0836