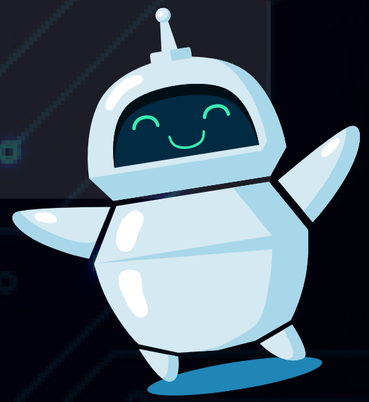
WHISPER

Boosting the model



Makarov Igor
Sokolov Pavel
Alshevskiy Dmitriy

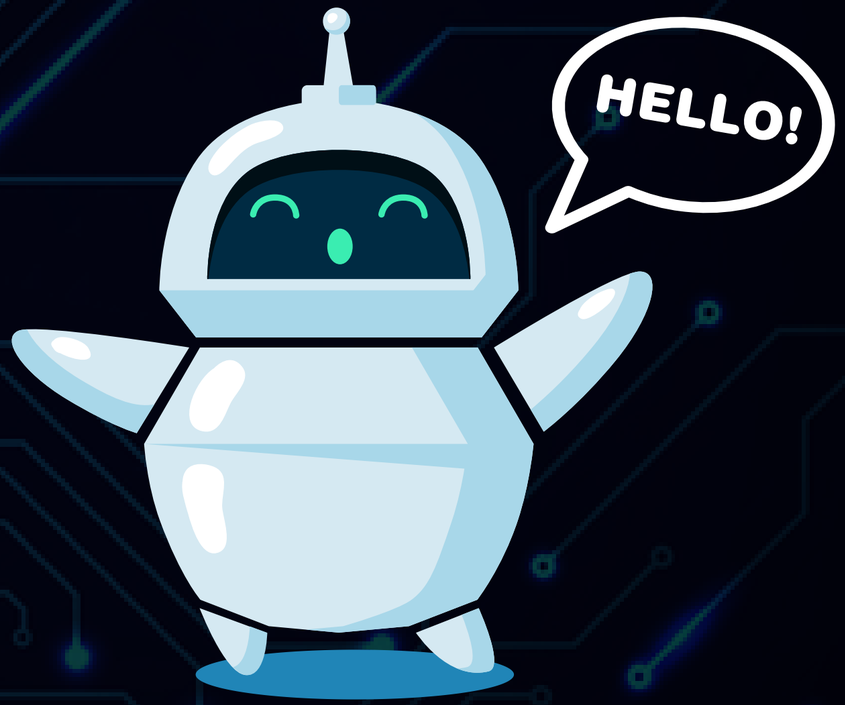
INTRODUCTION

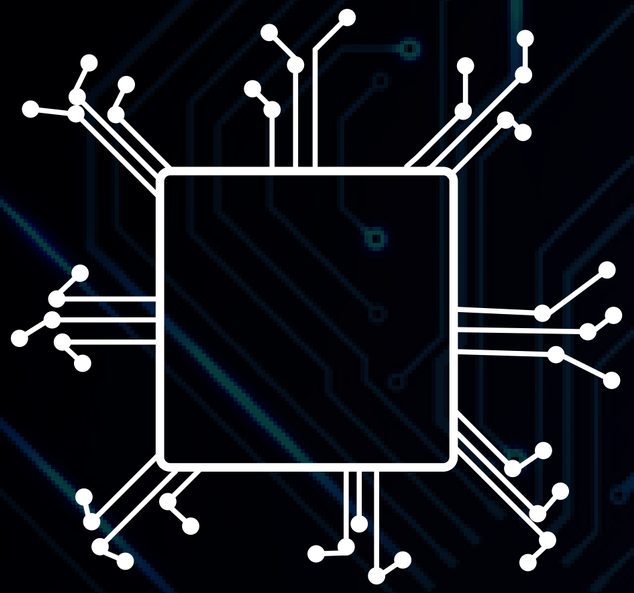
```
WhisperForConditionalGeneration(
  (model): WhisperModel(
    (encoder): WhisperEncoder(
     (conv1): Conv1d(128, 1280, kernel_size=(3,), stride=(1,), padding=(1,))
       (conv2): Conv1d(1280, 1280, kernel_size=(3,), stride=(2,), padding=(1,))
            (embed_positions): Embedding(1500, 1280)
            (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

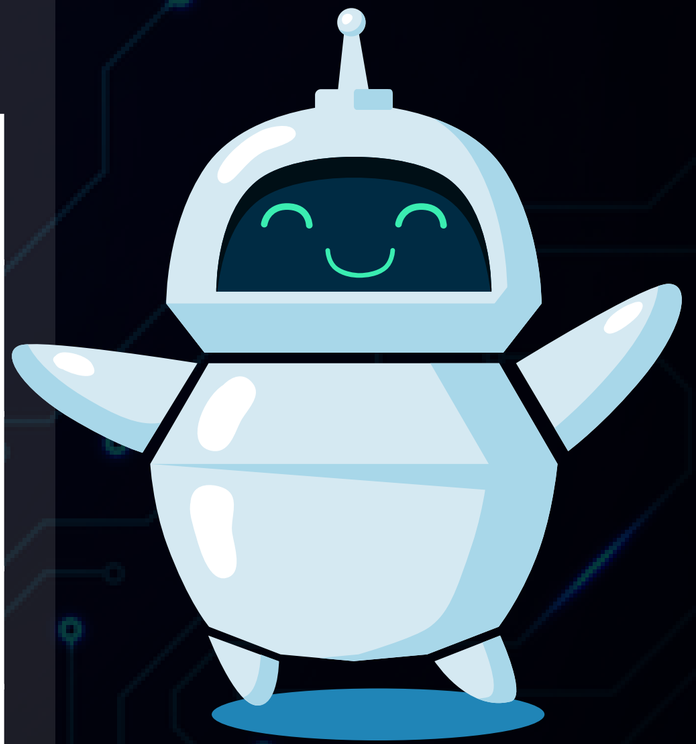
CPU:
aten::addmm типа output = beta * input + alpha * (mat1 @ mat2)

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.
- L2 прунинг. Разрушил модель при проходе, решено не использовать.
- Варианты прунинга с удалением весов реализовать не удалось.

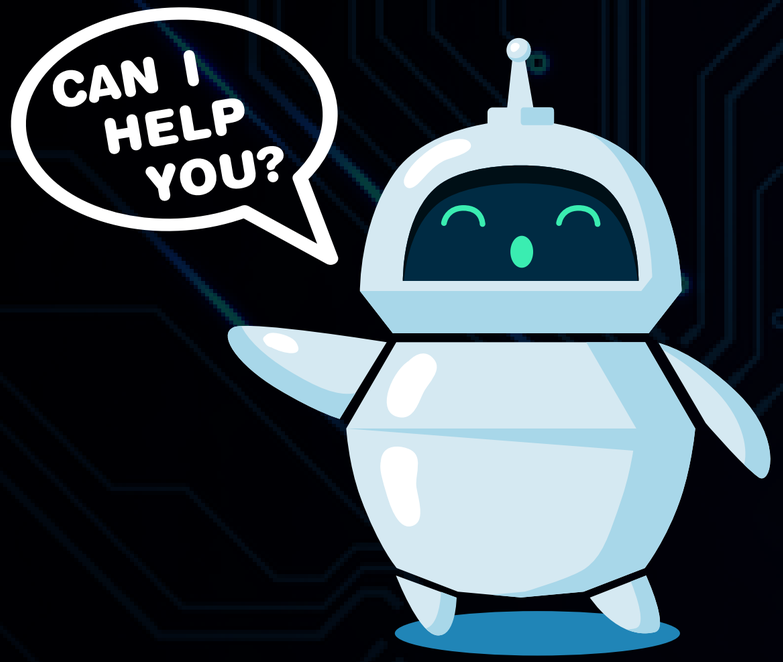
Pruning rate	WER	dWER	CER	dCER	Avg Time per Audio (s)	Delta
GPU base	0.4403	-	0.1583	-	0.8386	-
GPU 0.81	0.4410	– 0.16%	0.1619	– 2.22%	0.6089	– 27.39%
CPU base	0.4403	-	0.1583	-	8.7137	-
CPU 0.81	0.4410	– 0.16%	0.1619	– 2.22%	7.4373	– 14.65%



QUANTIZATION BASELINE

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

Quant	WER	dWER	CER	dCER	Avg Time per Audio (s)	Delta
GPU base	0.4403	-	0.1583	-	0.8386	-
GPU q	0.4472	– 1.54%	0.1586	– 0.19%	3.0069	+ 258.56%
CPU base	0.4403	-	0.1583	-	8.7137	-
CPU q	0.4740	– 7.11%	0.1664	– 4.86%	5.8057	– 33.37 %



DYNAMIC QUANTIZATION CONFIGS

Quant	WER	dWER	CER	dCER	Avg Time per Audio (s)	Delta
CPU base	0.4403	-	0.1583	-	5.4567	-
Int8AInt8W	0.4480	– 1.71%	0.1582	+ 0.06%	4.8465	– 11.18%
Int8AInt4W	0.4498	– 2.12%	0.1610	– 1.68%	10.5658	+ 90.63%

FINAL RESULTS

Configuration	WER	dWER	CER	dCER	Avg Time per Audio (s)	Delta
GPU base	0.4403	-	0.1583	-	0.8386	-
CPU base	0.4403	-	0.1583	-	8.7137	-
CPU q	0.4740	– 7.11%	0.1664	– 4.86%	5.8057	– 33.37 %
CPU p 0.81	0.4410	– 0.16%	0.1619	– 2.22%	7.4373	– 14.65%
CPU q+p+c	**0.4815**	**– 8.56%**	**0.1664**	**– 4.86%**	**5.0836**	**– 41.66 %**

Real Time Ratio ~ 80.59%