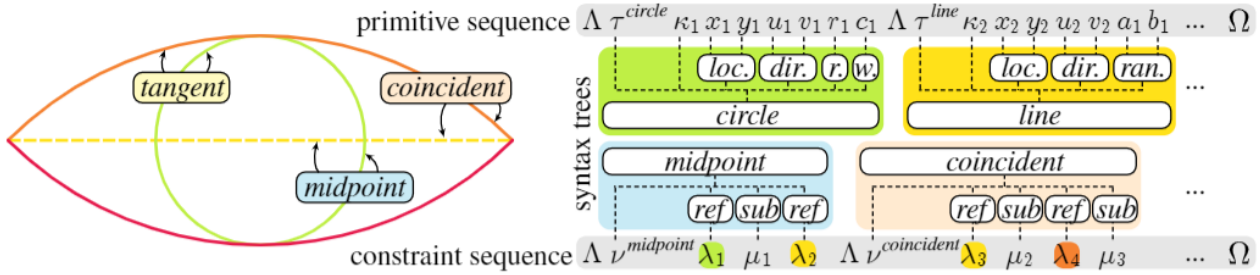
Training Transformers

Wamiq Reyaz
KAUST
16-Sep-2021

What do I train?

- Generative models for structured sequence generation.



What do I train?

- Training loss is Cross-Entropy.
- Generation is autoregressive.
- Both Decoder and Encoder-Decoder style architectures.
- GPT-2 style attention layers.
- 16-22 layers.
- Adam/AdamW optimizer
- lr=10e-4 seems to work well for me.

KAUST

What do I train?

```
config = GPT2Config(
        vocab_size=args.vocab,
        n_positions=args.enc_n,
        n_ctx=args.enc_n,
        n_embd=args.dim,
        n_layer=args.enc_layer,
        n_head=args.n_heads,
        is_causal=True,
        is_encoder=False,
        n_types=args.n_types,
        n_stypes=args.n_stypes
        )
```

KAUST

What do I train?

```
config = GPT2Config(
        vocab_size=2**8,
        n_positions=200,
        n_ctx=200,
        n_embd=544,
        n_layer=22,
        n_head=12,
        is_causal=True,
        is_encoder=False,
        n_types=args.n_types,
        n_stypes=args.n_stypes
        )
```

KAUST

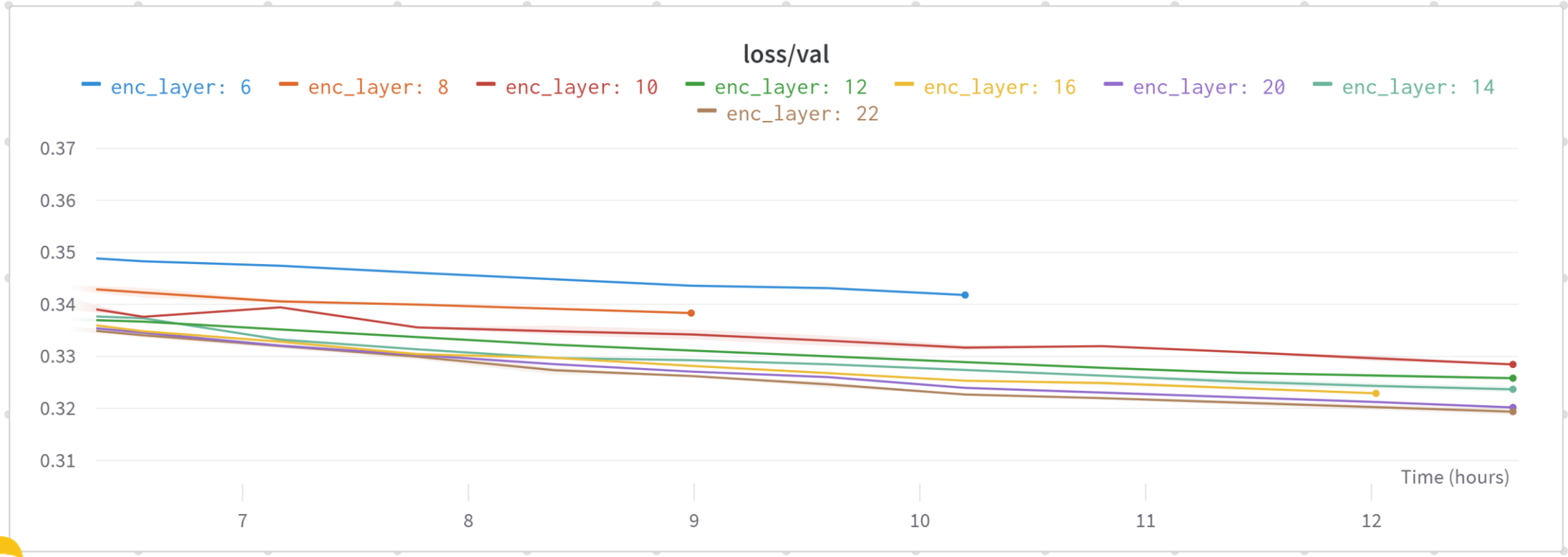
Where do I train it?

- 8x V100 or 8x A100 nodes.
- Train with DistributedDataParallel.
- Train with Mixed Precision (AMP).

KAUST

Lessons Learned

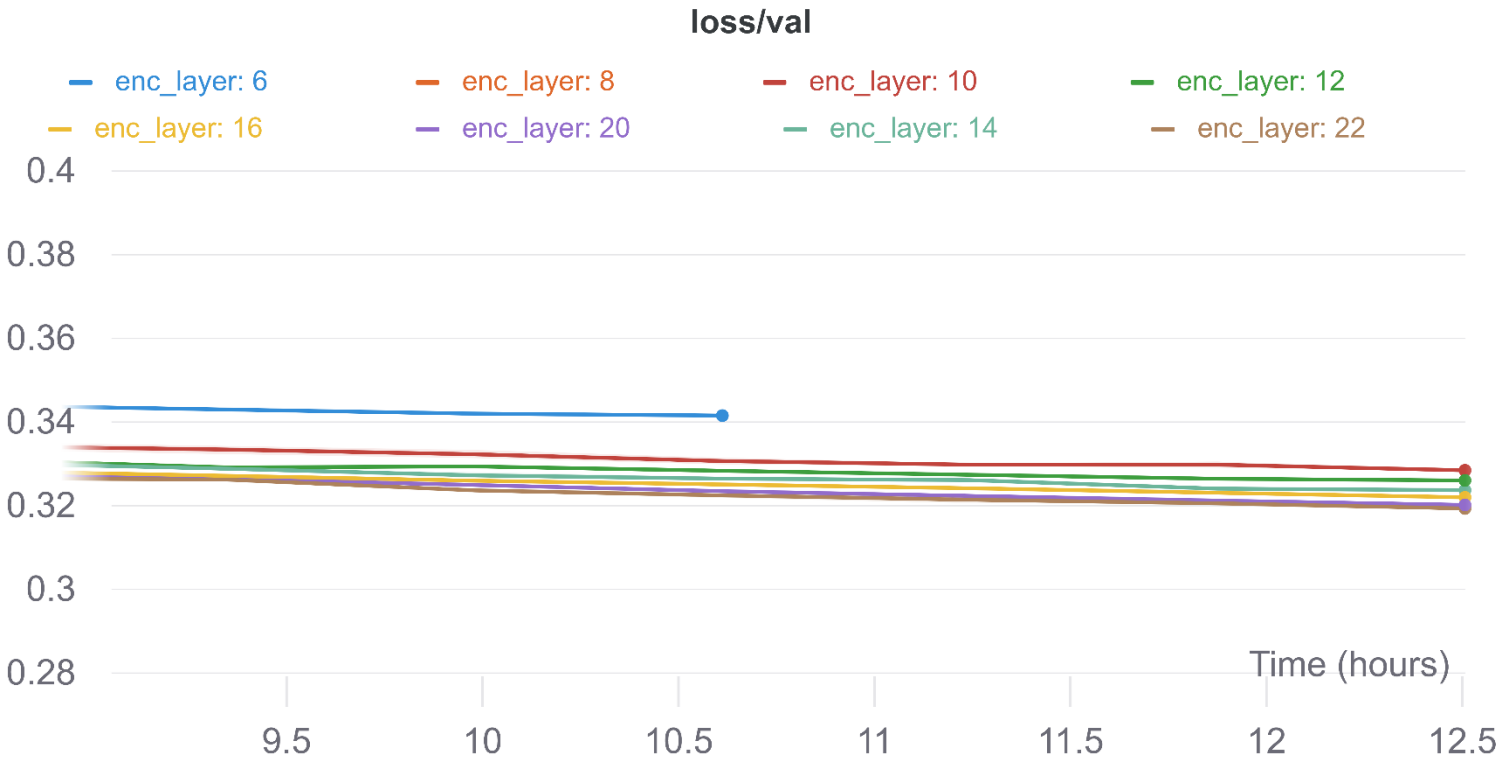
- Use larger models if your data is large.



KAUST

Lessons Learned

- Large in Layers



KAUST

Lessons Learned

- Large in Dimension

KAUST

Lessons Learned

- Use larger models if your data is large.
- Use DDP (about 10-25% [1] faster).
- Use AMP (saves about 25% [1] memory).
- Use Gradient Clipping to counteract divergence.
- Learning rate did not make a lot of difference in my experience.

KAUST

[1] Very rough numbers

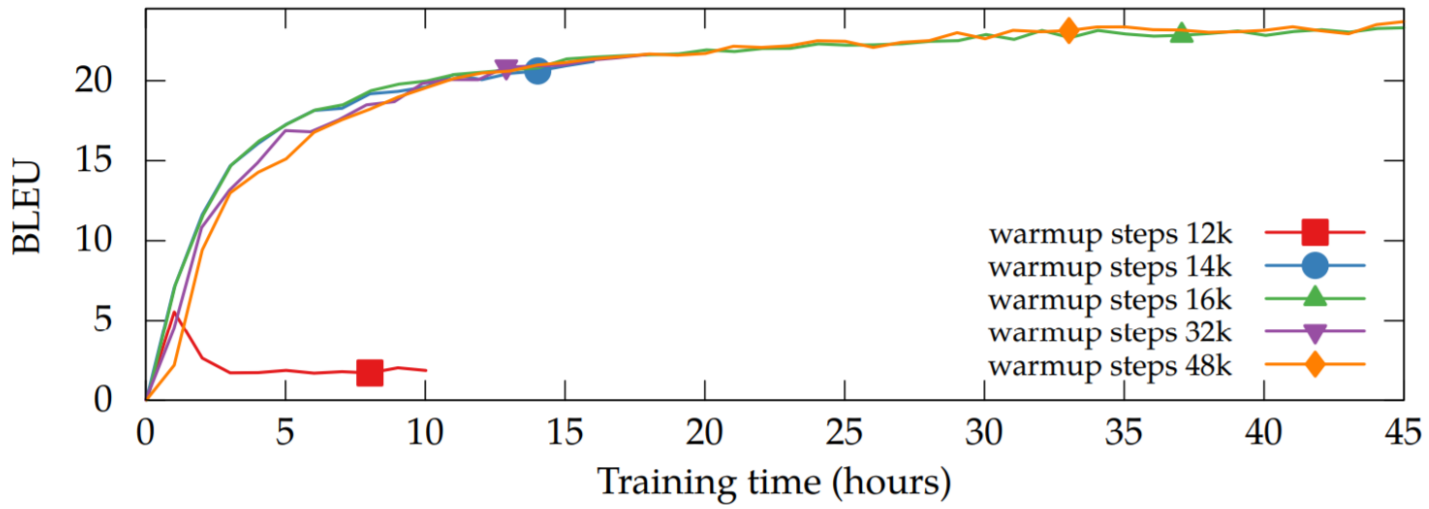
Lessons Learned (Common Problems)

- This might be specific to generative models.
- Use lower temperature for sampling. Higher temperatures lead to a lot of unparseable samples.
- If output samples do not make sense, make sure start tokens/conditional tokens are the same as during training. Can mess up very easily.
- Can perform masking of logits during sampling. Very time-consuming to implement. But yields benefits.

KAUST

Information Acquired

- Warmup improves performance somewhat.



KAUST

Figure from https://arxiv.org/abs/1804.00247

Information Acquired

- Batch size improves performance.

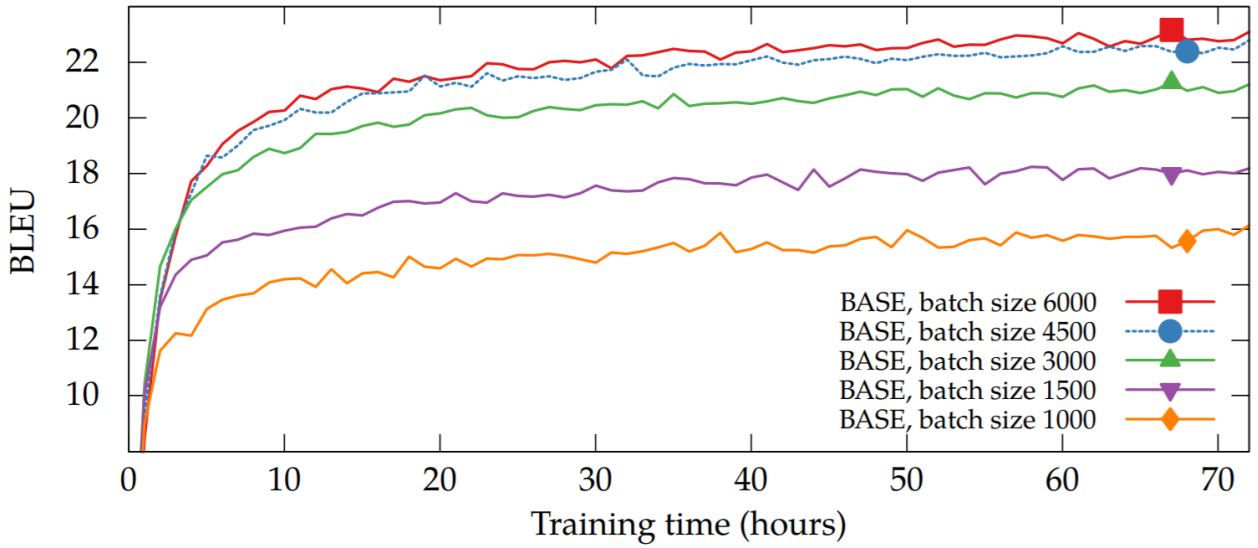


Figure from https://arxiv.org/abs/1804.00247

Final Takeaways

- Use larger models if your data is large.
- Use DDP (about 10-25% [1] faster).
- Use AMP (saves about 25% [1] memory).
- Use Gradient Clipping to counteract divergence.
- Use Warmup.
- Use max bs that fits on your GPUs.

KAUST

[1] Very rough numbers

References

- **Training Tips for the Transformer Model,** https://arxiv.org/abs/1804.00247

KAUST