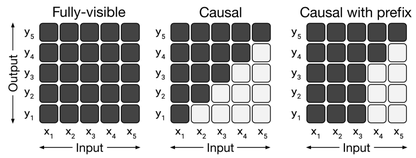
Going Beyond the Causal Mask in Language Modeling

Franz Srambical
p(doom) & Helmholtz Munich

A primer on the causal mask

For causal LMs at inference time, even if n tokens have already been generated, token k with $k < n$ cannot attend to token j with $k < j < n$, even though token j is already known (Figure from Raffel, Colin et al. 2020).



- Ideally, we want to omit the mask at inference time
- Omitting the mask on a causally pre-trained LM leads to distribution shift

Addressing common misconceptions

- The causal mask is not needed to prevent information leakage from future tokens during training
- The causal mask is not needed for parallel training
- The causal mask is not needed for teacher-forcing

You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.
You Don't Need the Causal Mask for LMs.

Right??

p(doom)

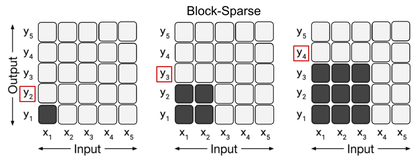
Going beyond the causal mask

- In PrefixLMs, full attention is used in the prefix, causal attention is used in the target
- Prefix language modeling leads to sparse supervision in unlabeled regime (prefix is not predicted)
- **Block-Sparse Language Modeling (BS-LM)**: Take all previous tokens as 'prefix' and each next token as sole 'target'

BS-LM simulates the mask-free inference regime.

You *do* need the causal mask for LMs

During BS-LM training, the mask depends on the position of the token to be predicted, i.e. a naïve implementation would require `seq_len` times as much memory and computation compared to causal LMs since the computational path splits token-wise after the first attention map calculation.



Moving beyond the causal mask necessitates mitigating the memory and compute overhead that comes with that.

The causal mask is a feature, not a bug.

twitter anon, 2024