Examining Scaling and Transfer of Language Model Architectures for Machine Translation

Biao Zhang, Behrooz Ghorbani, Ankur Bapna, Yong Cheng, Xavier Garcia, Jonathan Shen, Orhan Firat

Google Research

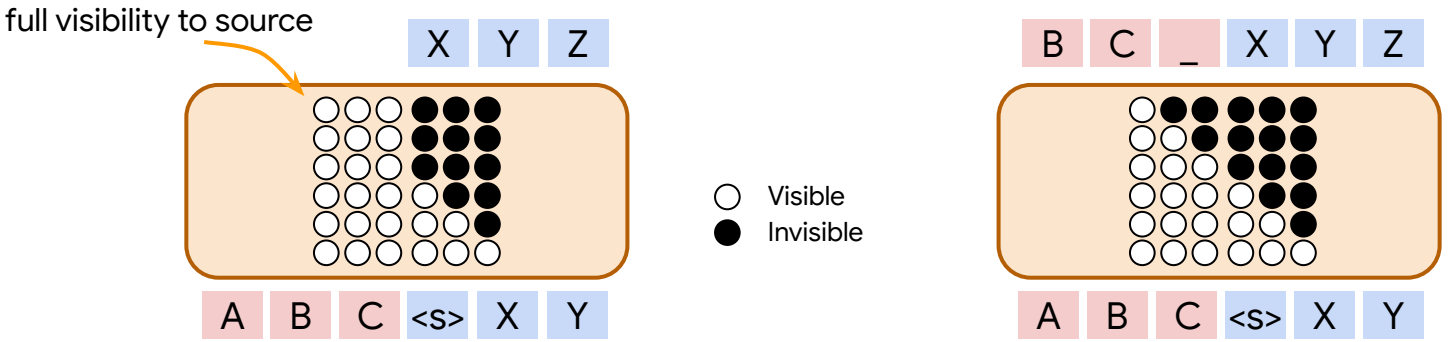
The Thirty-ninth International Conference on Machine Learning (ICML 2022)

Why Language Models for Translation?

- Language models have shown great performance with large-scale pretraining, and enable in-context learning
- Language models encode different inductive biases compared to encoder-decoder models, which might benefit translation
- However, how language models work for translation has been rarely studied
- We explore this question **jointly with model scaling and cross-lingual transfer**

Brown et al., 2020; Raffel et al., 2020; Xue et al., 2021; Wang et al., 2021

Language Model Architectures for Translation



PrefixLM

- Bidirectional attention over source input
- Only target-side induced MLE loss

CasualLM:

- Strict (causal) language model
- Both source + target MLE loss

Using one module to jointly perform **understanding** and **generation**

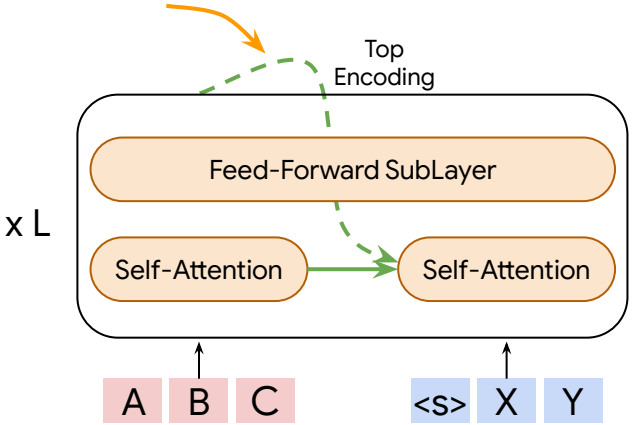
Model Variants: Examining More Design Choices

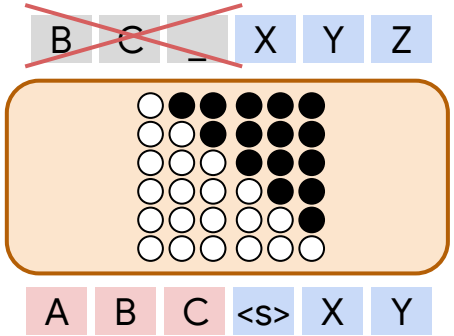
PrefixLM Using Top Layer Encodings (TopOnly)

final-layer encodings for target attention

CasualLM Target Only Loss (TgtOnly)

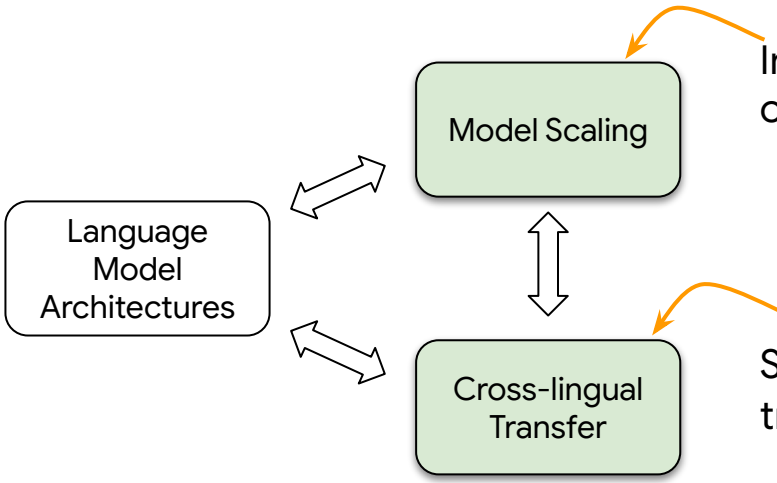
remove the source side training objective





Question: different LMs have different inductive biases, do they matter for translation?

On Model Scaling and Cross-lingual Transfer



Increasing the # model parameters has different effects on different models: **how does it affect LMs?**

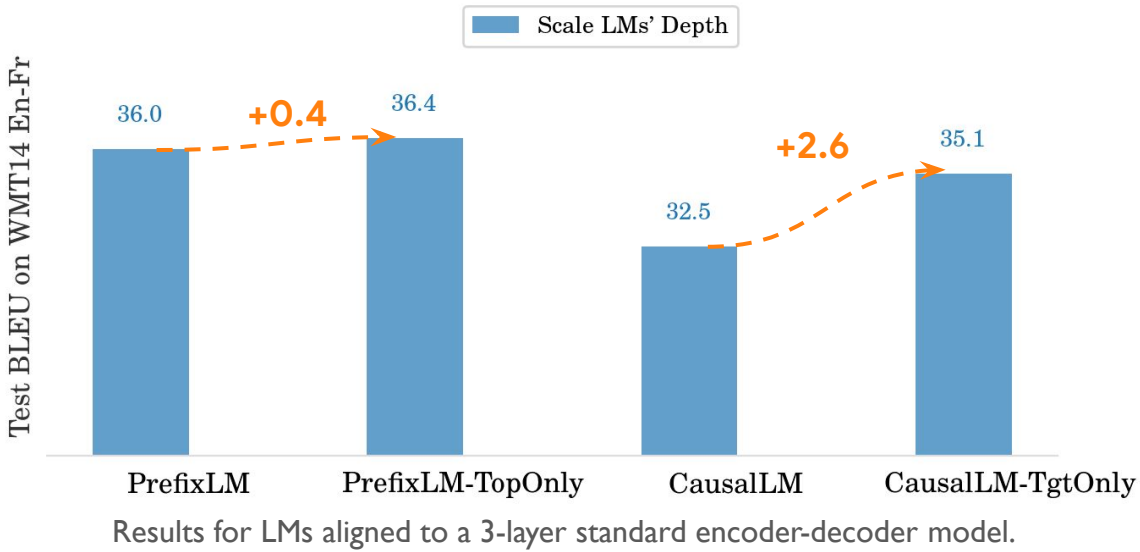
Sharing params over source & target (tying) affects the transfer dynamics: **how LMs behaves on transfer?**

We study the interaction of these three aspects of language models for translation

Experiments

- Transformer base & big model: 8/16 heads, 512/1024 model size
- Dataset
  - Bilingual: **WMT14 En-Fr**, WMT19 En-Zh, Web En-De (2B samples)
  - Multilingual: **WMT En-De/Zh/Fr**, OPUS-100 (Zhang et al., 2020)
- Model Scaling
  - Encoder-decoder: increase model depth
  - Language model: increasing either model depth ("**-Deep**") or model width ("**-Wide**")
- Evaluation
  - SacreBLEU
  - Log-perplexity score (PPL) for scaling

Do Design Choices Matter? **Yes!** Especially for Small-size Model

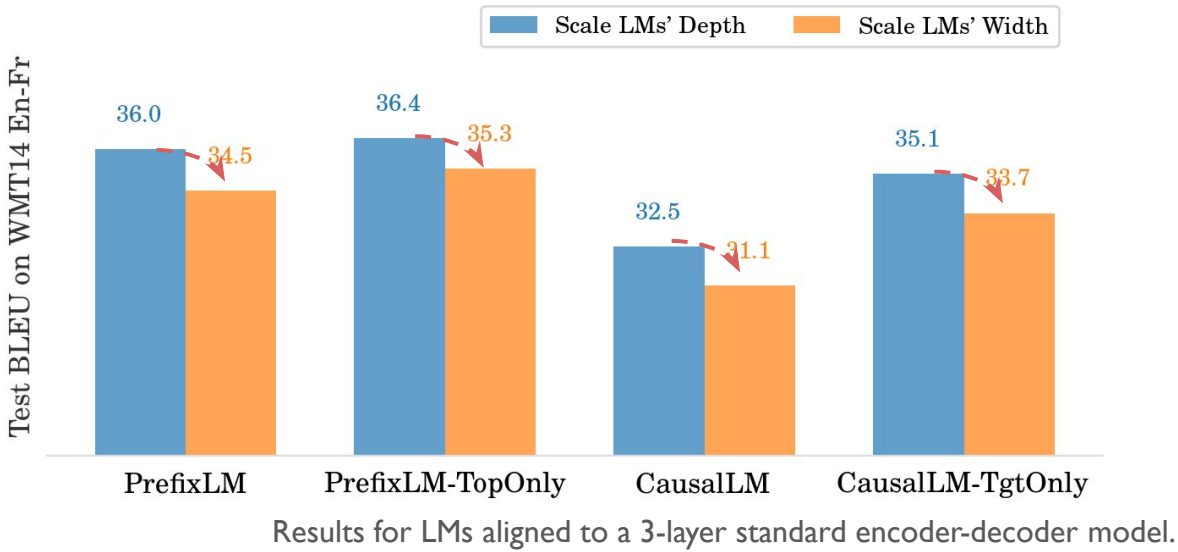


Results for LMs aligned to a 3-layer standard encoder-decoder model.

PrefixLM > CausalLM

- **PrefixLM**: Using final-layer source encodings work better for translation
- **CausalLM**: Adding the source-side training objective doesn't improve quality

Do Design Choices Matter? **Yes!** Especially for Small-size Model

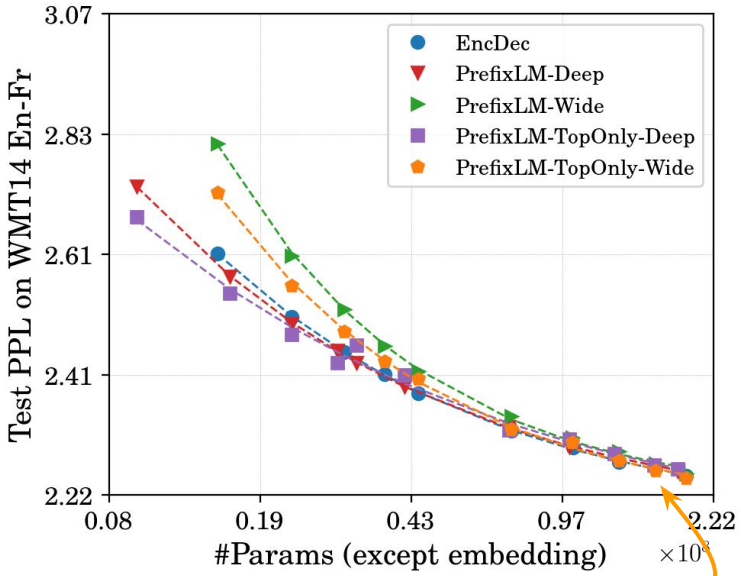


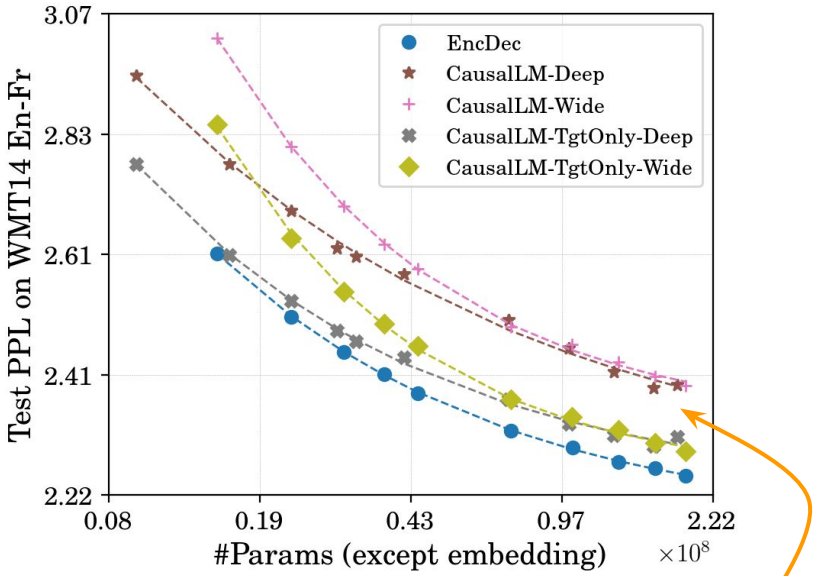
Results for LMs aligned to a 3-layer standard encoder-decoder model.

Deep > Wide

Increasing depth is more effective for language modeling than increasing width

Does Model Scaling Matter? **Yes**! Gap narrows at scale

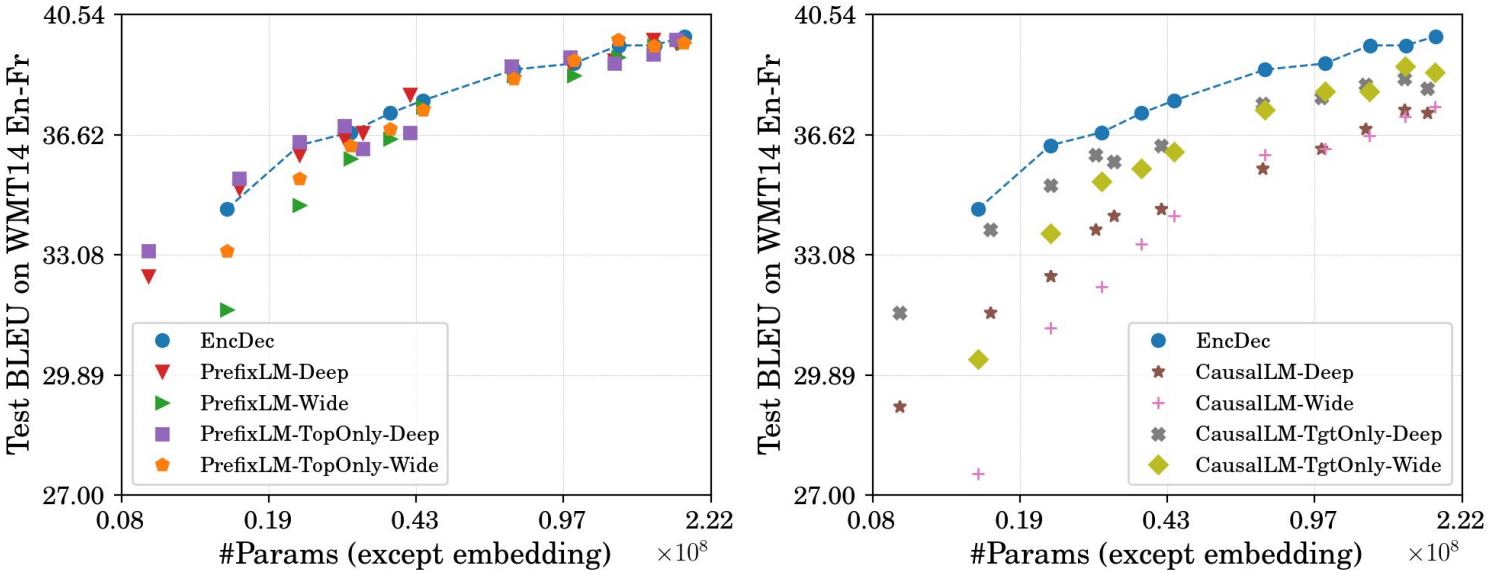




PrefixLM and EncDec, deep and wide models converge to similar bands.

CausalLM still retains a gap, especially with the source loss.

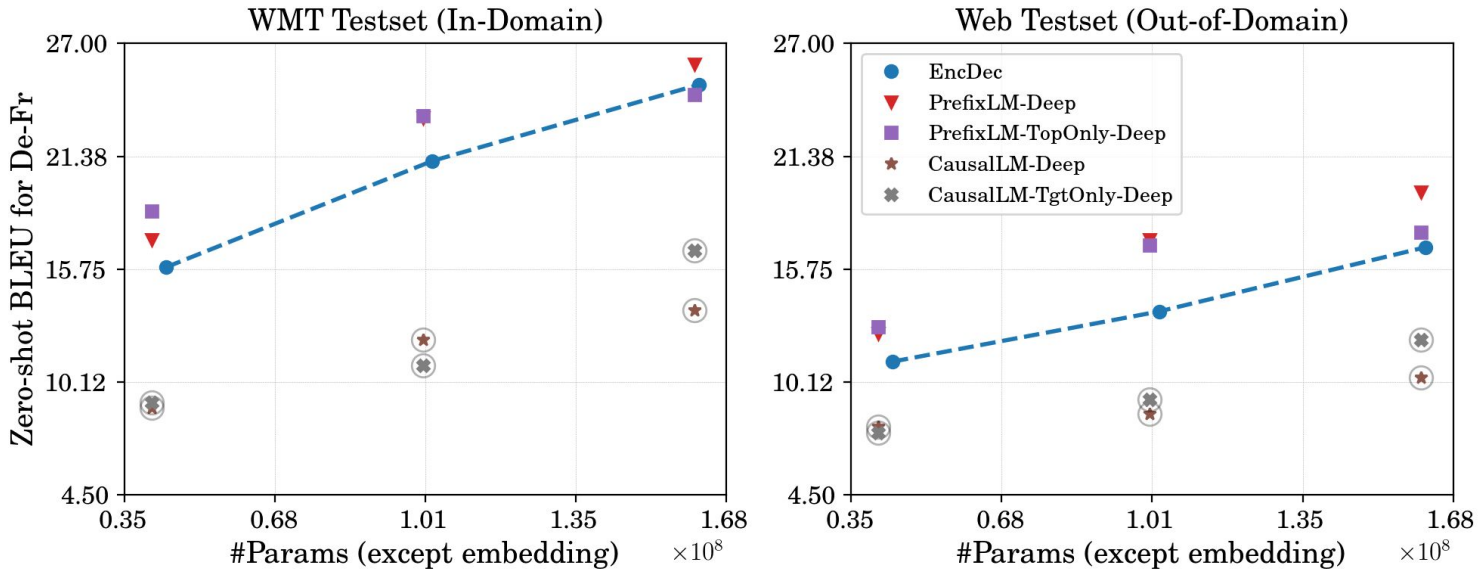
Does Model Scaling Matter? **Yes**! Gap narrows at scale



BLEU scores show similar trends. Still, LMs tend to underperform EncDec.

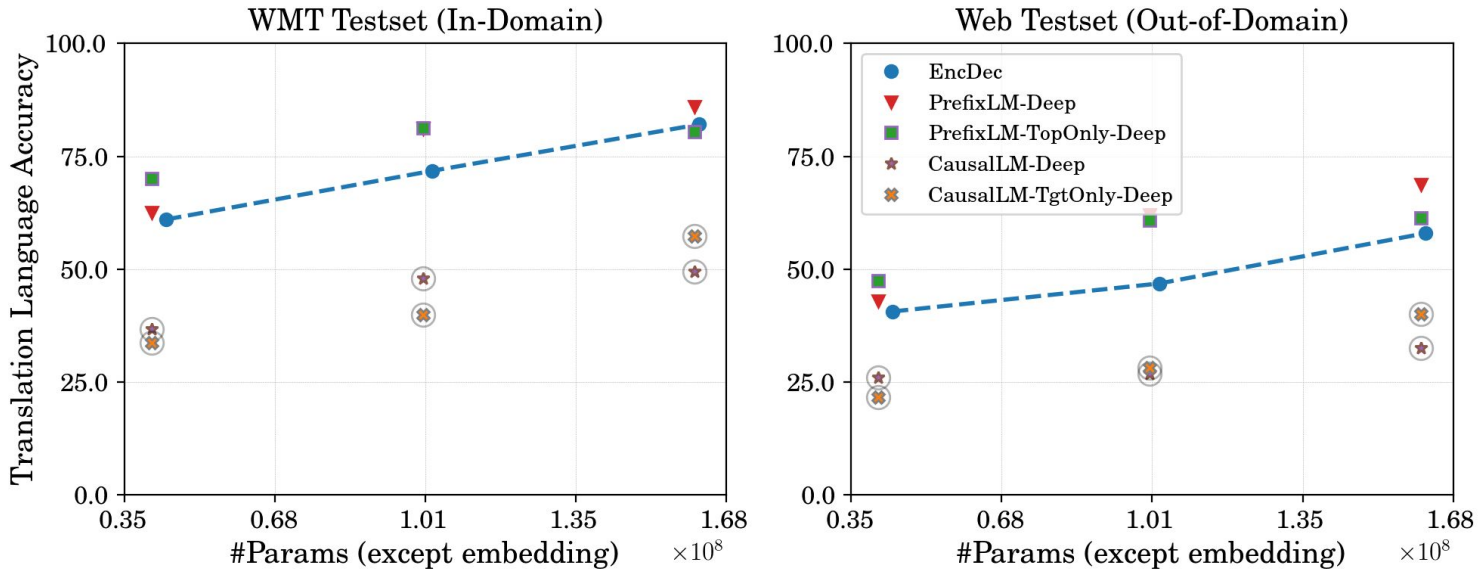
Note the relationship between BLEU and PPL is non-trivial (Ghorbani et al., 2021)

How LMs and Scaling Affect Cross-lingual Transfer?



- ✔ Model scaling improves cross-lingual transfer for all models
- ✔ PrefixLM greatly improves zero-shot translation.

How LMs and Scaling Affect Cross-lingual Transfer?



The improvement of PrefixLM comes from its reduction of off-target translation (Zhang et al., 2020)

To Summarize

- Language model architecture matters for translation
  - PrefixLM > CausalLM, Deep > Wide, TopOnly > Layerwise, TgtOnly > Src+Tgt
- Model scaling matters a lot
  - The impact of architectural differences gradually reduce as models are scaled up
  - The whole scaling picture is recommended for model comparison in the future
- Surprising impact on cross-lingual transfer
  - PrefixLM largely benefits zero-shot transfer

Paper: https://arxiv.org/abs/2202.00528