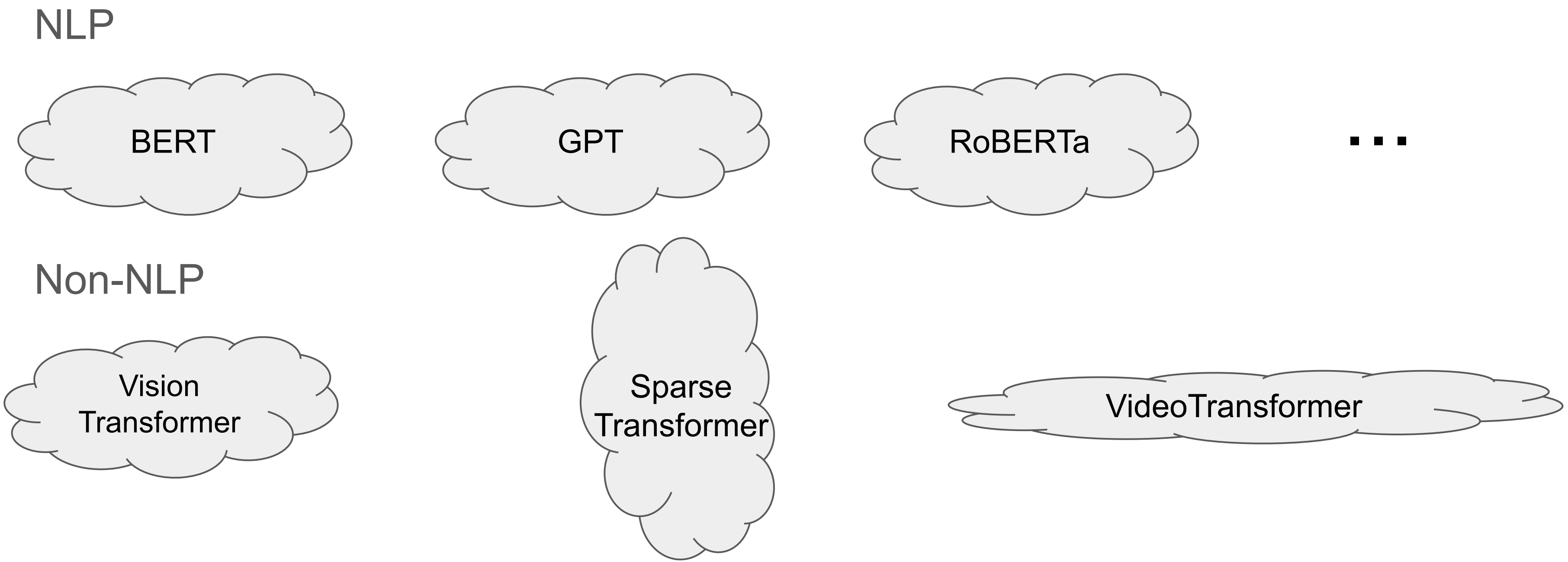
Which transformer architecture fits my data?
A vocabulary bottleneck in self-attention

Noam Wies, Yoav Levine, Daniel Jannai, and Amnon Shashua
The Hebrew University of Jerusalem

Transformers across domains



NLP

BERT

GPT

RoBERTa

...

Non-NLP

Vision Transformer

Sparse Transformer

VideoTransformer

Depth-to-width ratio varies across applications

Depth-to-width ratio

Henighan et al. 2020:

Modality	Optimal Depth-to-width ratio
Text	1/50
Images	1/10
Math	1/5

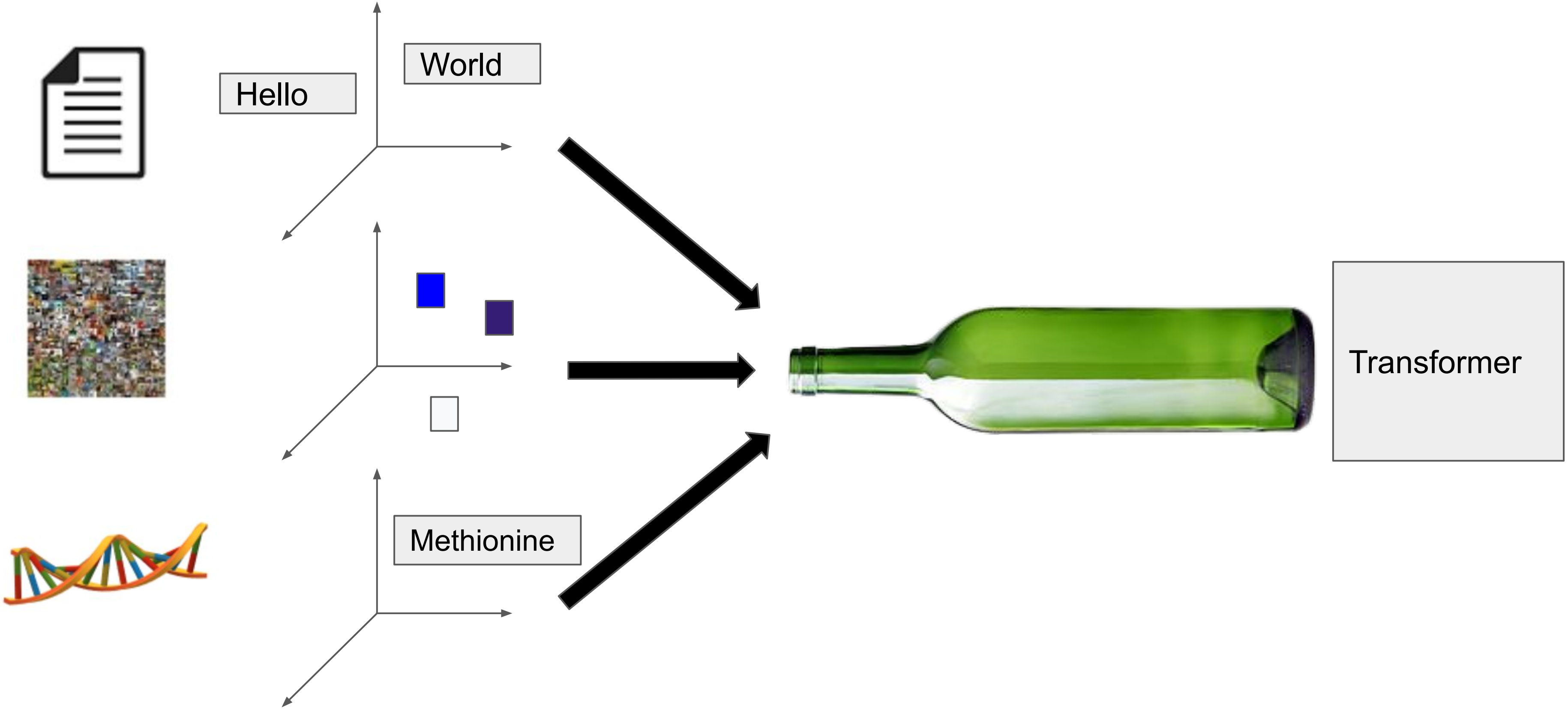
Subtleties, *e.g.*, in vision:

Vision Transformer

Sparse Transformer

Levine et al. 2020:
*“From an **architecture expressivity** perspective, each Transformer size has an optimal depth.”*

Input Embedding Layer

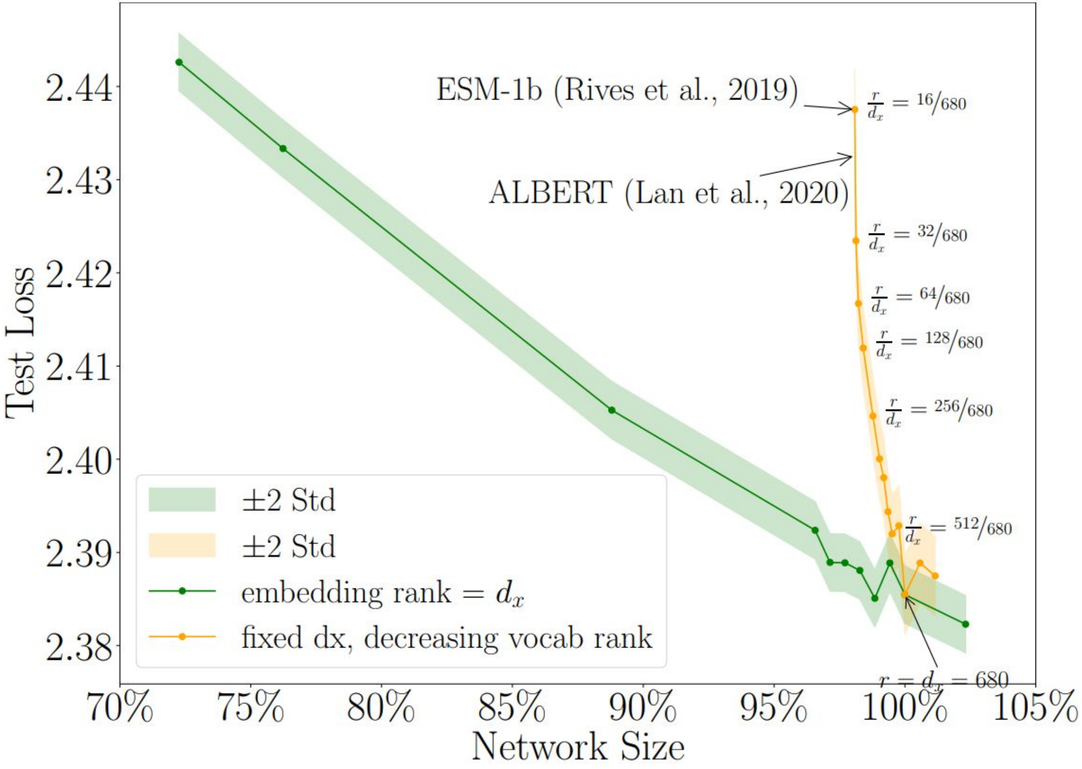


The Vocabulary Bottleneck

Informal theorem:
For any data modality, the capacity of Transformer to model inputs dependencies scale like
min{d ,rank (V)}.

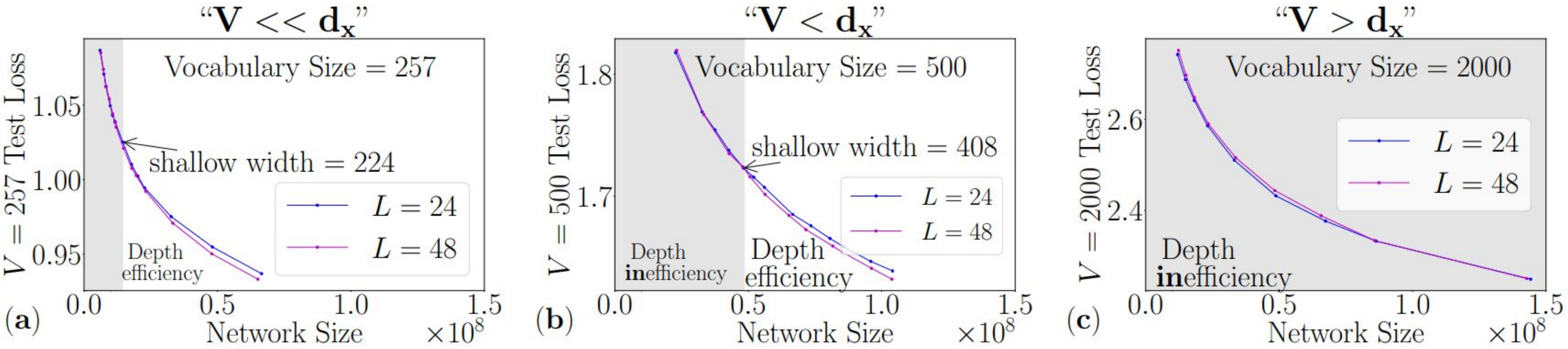
width

Input embedding



Vocabulary affects the depth-to-width interplay

small vocabulary => deeper is better earlier



Domain-independent guidelines for Transformer architecture design!