Google Transformer Sizes

GPipe (Huang et al., 2019) introduces microbatches for faster training of deep models across multiple GPUs.

Enc/Dec Depth	FF Dim	Heads	Total Parameters	GPUs Used	
6	8192	16	400M	1 default	
12	16384	32	1.3B	2	“wide”
24	8192	16	1.3B	4	“deep”
32	16384	32	3.0B	8	
64	16384	32	6.0B	16	

- “Deep” better than “wide” on low-resource languages.
  - Indicates better generalization.
- Further tricks needed to keep the training stable.