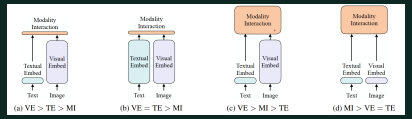
ViLT

VLP (vision-and-language model)目前的四种方式，ViLT使用了较清亮的vision 和 text编码方案



3种pre-train task (具体做法就是在transformer抽取的特征后加特定的任务头)

- Image Text Matching
- Masked Language Model
- Word Patch Alignment

Whole Word Masking