# Attention Is All You Need

https://www.bilibili.com/read/cv13759416/

* + 2017 NeuralIPS
  + Google Research
  + Dataset: WMT 2014 Englishto-German/French translation
  + 法：a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.
  + Optimizer:
  + 设备：某个任务3.5天，8GPUs；某个任务： twelve hours on eight P100 GPUs
  + 优：
    1. **纯RNN**：RNN产生hidden state(记住上一时刻的hidden state+input)，所以天然表达了input的顺序，即位置信息，但是无法并行（因为上一时刻没算粗来，这一时刻算不出来）； 有用factorization因式分解、有条件计算的方法
    2. **RNN+attention**
    3. **CNN+attention**
    4. **纯attention**：本文
  + 缺：
  + 代码：https://github.com/tensorflow/tensor2tensor

Tensorflow 早在2011年，Google Brain（谷歌大脑）内部孵化了一个叫做DistBelief的项目，这个项目也被视为TensorFlow的前身；2015年11月 ，Google发布了TensorFlow的白皮书并很快将其开源。但即使以Google的影响力，开源TensorFlow并没有在技术圈之外引起过多反响

PyTorch是2017年1月FAIR(Facbook AI Reesearch)发布的一款深度学习框架

* + 未来：
    1. 用于解决不仅文本input/output，但是audio.image.video.轨迹这些维数特别大，需要trick
    2. 用multi-head attention 每个head求attention方法不同：我全都要，比如random/restricted/stride/global求attention，排列组合