# Transformer：一次注意力怎么算？

建议 35-45 分钟；先用纸笔预测，再由老师投影实验核对。本地模型课需要电脑运行下载的代码。进阶运算可分两节课。

## 学习目标

区分 query、key、value，计算三个权重的和，并解释因果遮罩为什么不能偷看未来。

## 先修与准备

先学猜词和点积；softmax 可先看比例图，再由大孩子计算。

## 材料

新学习单、计算器可选；投影的两维注意力实验。

## 课堂顺序（5 / 10 / 10 / 10 分钟，可另留 5-10 分钟讨论）

1. 先猜：query=[1,0]，keys=[1,0]、[0,1]、[1,1]。先不算指数，比较哪两个点积一样。

2. 动手：分数除以 √2，再做 softmax：每个分数取指数，然后除以指数总和。三个权重约为 0.401、0.198、0.401。

3. 观察：values=[1,0]、[0,1]、[2,1]，用同一权重分别混合每一维，得到约 [1.203,0.599]。权重不是预测词的概率。

4. 挑战：预测第 2 个位置之后的下一个词时，启用因果遮罩，不能使用第 3 个位置的信息。再把 query 换成 [0,1] 比较。

# Transformer：一次注意力怎么算？

## 做好的例子

query=[1,0] 与三个 key 的点积为 1、0、1；缩放后为 0.707、0、0.707。softmax 不是把原分数直接除以总和。

## 理解检查：先让学生独立回答

遮住第三个位置后，它的权重是多少？能仅凭注意力最高的词证明模型为什么给出某个答案吗？

## 参考答案与学习单核对

第三位权重 0；前两位重新归一化，约 0.670、0.330，输出约 [0.670,0.330]。不能把注意力热图当充分的因果解释。更换 query=[0,1] 时，未遮罩权重约 0.198、0.401、0.401。 query=[0,1] 的点积为 0,1,1。三个 key 完全相同且无遮挡时，权重各为 1/3；序列模型仍需要位置信息。

## 不要把结论讲大了

数值是人为设置的单头示例，没有学到词义。完整 Transformer 还包含学习到的投影、多头、位置表示、前馈层、残差等；这里不训练完整模型。

## 教学调整

需要帮助时：先只做例子，由老师朗读题目，用实物或纸格代替公式。已经掌握时：一次只改一个条件，先记预测，再验证。遇到不一样的答案，先查规则、编码或数据，不先判学生错。

## 带回家的活动

给家长三个已经算好的权重，检查和为 1，再手算混合后的 value。

## 接着学

猜词游戏：AI 怎么读懂一句话 / CNN：带着同一把尺子找横线 / 自己的小语言模型：从计数到生成

## 教师参考（完整链接也在课程页）

Vaswani et al.: Attention Is All You Need: https://arxiv.org/abs/1706.03762