Transformer: Calculate One Attention Step

Plan 35-45 minutes: predict on paper, then check on one projected computer. The local-model lesson needs a computer to run its download. Split advanced calculations across two sessions.

Learning goal

Distinguish queries, keys and values, check that weights sum to one and explain why causal masking hides the future.

Before starting

Start with Guess the Word and dot products; younger learners can inspect proportions before calculating softmax.

Materials

The new worksheet, an optional calculator and the projected two-dimensional attention lab.

Class sequence (5 / 10 / 10 / 10 min, plus 5-10 min discussion)

1. Predict: query=[1,0], keys=[1,0],[0,1],[1,1]. Before exponentials, identify the tied dot products.

2. Try: divide scores by √2, exponentiate each and divide by the total. Weights are about 0.401,0.198,0.401.

3. Observe: mix values [1,0],[0,1],[2,1] with those weights to get about [1.203,0.599]. These are not next-word probabilities.

4. Challenge: when predicting the next token after position two, enable the causal mask so position three is unavailable. Then compare query=[0,1].

Transformer: Calculate One Attention Step

Worked example

Dot products for query=[1,0] are 1,0,1; scaled scores are 0.707,0,0.707. Softmax does not simply divide raw scores by their sum.

Check understanding: ask before showing an answer

What is position three’s weight after masking it? Does the highest attention weight prove why a model answered?

Expected answers and worksheet check

Position three gets zero; the remaining weights renormalize to about 0.670,0.330, giving output [0.670,0.330]. An attention map is not a sufficient causal explanation. With query=[0,1] and no mask, weights are about 0.198,0.401,0.401. With query=[0,1], dot products are 0,1,1. Three identical keys without a mask give weights of 1/3 each; a sequence model still needs positional information.

Keep the claim within its limits

These are hand-set numbers for one attention head without learned word meanings. A complete Transformer also has learned projections, multiple heads, position information, feed-forward layers and residuals.

Adapt the lesson

For support, work through just the example, read prompts aloud and use counters or paper grids. For extension, change one condition at a time, write a prediction and test it. When answers differ, inspect rules, encoding and data before judging a child’s answer.

At home

Give a parent three computed weights, check they sum to one and calculate the mixed value.

Continue learning

Guess the Word: How AI Reads a Sentence / CNN: Find a Line with One Shared Ruler / Your Own Tiny Language Model: Count, Then Generate

Teacher reference (links also on the lesson page)

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