Scientific Computing in Python / 2027 Lecture 6

Embeddings and scaled dot-product attention

Lecture notes

Attention combines a set of value vectors using weights obtained from query-key comparisons. We shall calculate that operation explicitly, keep track of every matrix dimension, and derive its normalization and masking properties. The supplied example uses fixed invented coordinates. It illustrates the computation inside a larger learned system without treating the coordinates as semantic or scientific evidence.

1 Embedding and projection matrices

Slides for this section

Let $X \in \mathbb{R}^{n \times d}$ contain one embedding row per position. Three linear maps produce queries, keys, and values. Queries and keys must have the same feature dimension for dot products; values may have a different dimension. The example's embeddings and matrices are hand-set, not learned from a corpus.

$$Q = XW_Q \in \mathbb{R}^{n \times d_k}, \quad K = XW_K \in \mathbb{R}^{n \times d_k}, \quad V = XW_V \in \mathbb{R}^{n \times d_v}.$$

The score matrix QK^T has shape $n \times n$.

Row and column meanings

$$S_{ij} = \frac{q_i^\mathsf{T} k_j}{\sqrt{d_k}}.$$

Row i fixes the query. Column j identifies a key whose associated value may contribute to that query's output.

2 Rowwise softmax

Slides for this section

For each query, exponentiate the scores and divide by their sum over allowed keys. The resulting weights are nonnegative and sum to one. Softmax weights describe this computational normalization; they are not automatically calibrated probabilities of a scientific claim.

$$A_{ij} = \frac{e^{S_{ij}}}{\sum_\ell e^{S_{i\ell}}}, \qquad A_{ij} > 0, \quad \sum_j A_{ij} = 1.$$

Normalization is performed independently for each row.

Stable normalization

For any row constant c,

$$\frac{e^{s_j - c}}{\sum_\ell e^{s_\ell - c}} = \frac{e^{s_j}}{\sum_\ell e^{s_\ell}}.$$

Choose $c = \max_j s_j$ to avoid exponentiating large positive numbers.

3 Weighted values

Slides for this section

The output row is the weighted sum of value rows. Since the weights are nonnegative and sum to one, it lies in their convex hull. At fixed queries, keys, and mask, the operation is linear in the values, although the whole attention map is nonlinear in its inputs because the weights also depend on them.

$$Z = AV, \qquad z_i = \sum_j A_{ij} v_j.$$

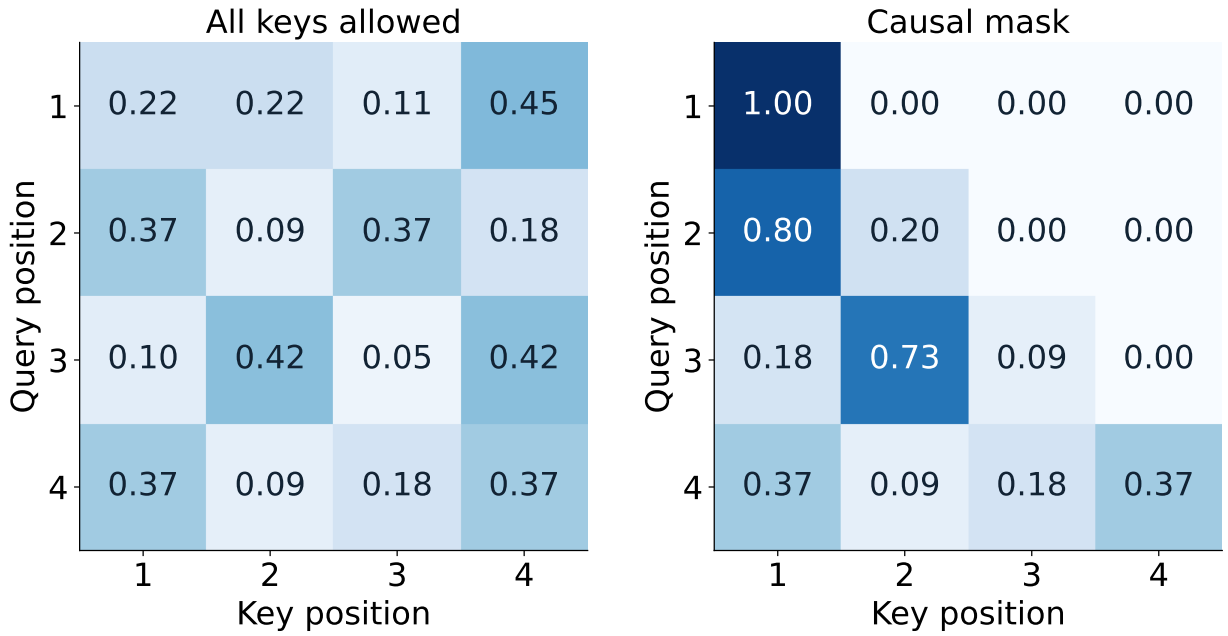
If every value is the same vector v_*, then $z_i = v_*$ for every row.

Scaling the values

With Q, K and the allowed-key set fixed,

$$A(\alpha V + \beta W) = \alpha AV + \beta AW.$$

Doubling V doubles Z. Changing Q generally changes the weights themselves.



The same four hand-set embeddings and projections yield different row-normalized weights when future positions are disallowed. Decimal labels are rounded; tests use full precision.

4 A two-key calculation

Slides for this section

Take a query $(1, 0)$ and keys $(1, 0)$ and $(0, 1)$, with $d_k = 2$. The scores are $1/\sqrt{2}$ and zero. Pair the keys with values $(2, 0)$ and $(0, 4)$. The output illustrates how a scalar score weight multiplies an entire value vector.

$$s = (1/\sqrt{2}, 0), \qquad A = \left(\frac{e^{1/\sqrt{2}}}{e^{1/\sqrt{2}} + 1}, \frac{1}{e^{1/\sqrt{2}} + 1} \right).$$

$$A \approx (0.66976, 0.33024).$$

The output vector

$$z = 0.66976(2, 0) + 0.33024(0, 4) \approx (1.33952, 1.32096).$$

The two output coordinates are value coordinates. They need not sum to one and are not token probabilities.

5 Causal masking

Slides for this section

A causal mask allows position i to use keys only at positions $j \leq i$. Masked scores are excluded before normalization. Every row must retain at least one allowed key; an entirely masked row has no defined normalization. In a single attention layer with rowwise projections, changing a future embedding cannot alter earlier causal outputs.

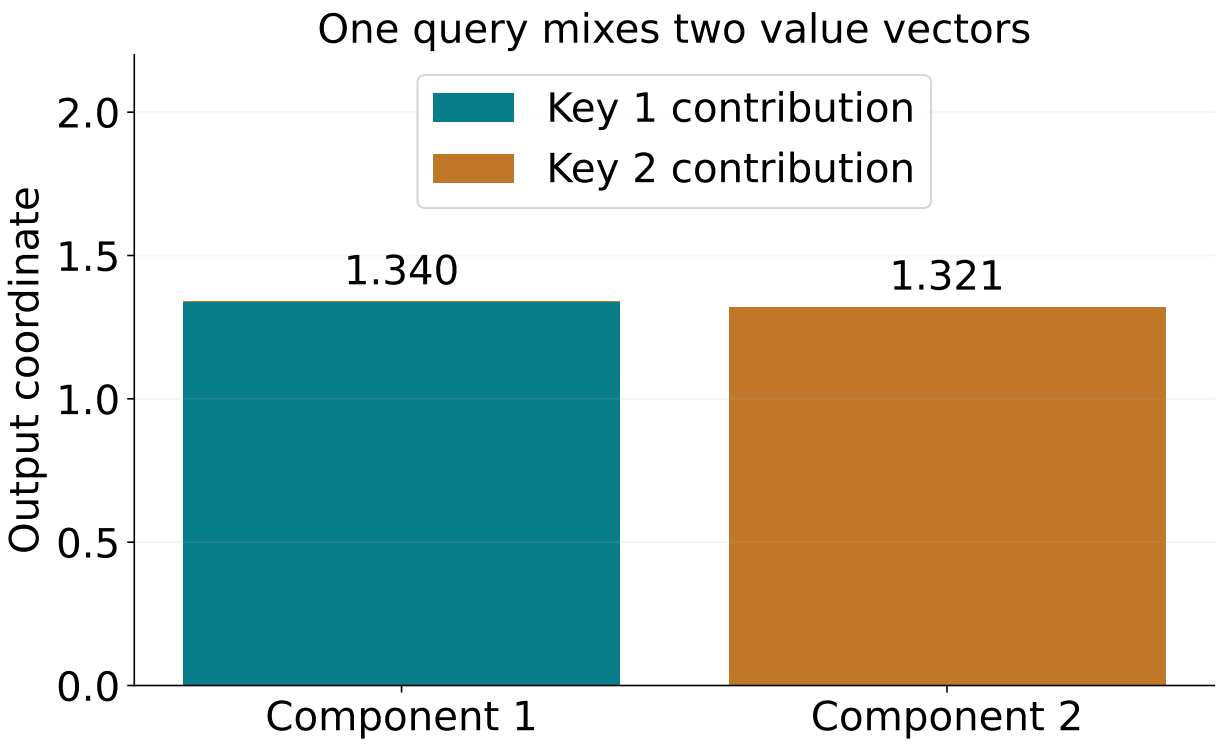
$$A_{ij} = 0 \quad (j > i), \qquad \sum_{j \leq i} A_{ij} = 1.$$

The first position has only one allowed key, hence $A_{11} = 1$.

A causal perturbation test

Change only the final embedding row while holding projection matrices fixed.

Earlier causal outputs remain unchanged. Earlier unmasked outputs can change because the final key and value are available to them.



For the worked query and two keys, weighted contributions sum to [1.33952, 1.32095]. These invented output coordinates are not token probabilities or truth scores.

6 Scaling and learned representations

Slides for this section

If independent query and key coordinates have mean zero and variance one, their dot product has variance d_k. Division by $\sqrt{d_k}$ then gives a score with variance of order one. This motivates the scale under those assumptions. A transformer combines attention with additional operations and learns its parameters from a training objective; the fixed example establishes none of that learned behavior.

Under independent unit-variance coordinates,

$$\operatorname{Var}\left(\sum_{r=1}^{d_k} q_r k_r\right) = d_k.$$

Dividing by $\sqrt{d_k}$ keeps the variance at one.

Permutation and position

Without positional information, permuting input rows by P gives

$$Q' = PQ,\ K' = PK,\ V' = PV \quad \Longrightarrow \quad Z' = PZ$$

for unmasked attention.

Position encodings or structured masks provide information not present in an unordered collection of embedding rows.

Exercises

P27-06-E1. Trace a new set of shapes

There are 5 token positions, embedding width 4, query/key width 3 and value width 2.
a Give the shapes of X, WQ, WK and WV.

b Give the shapes of Q, K, V, the score table and Z.

c A proposed implementation multiplies Q@K instead of Q@K.T. Explain the error.

P27-06-E2. A transparent three-value mixture

One query has equal scaled scores for three keys. Value rows are [3,0], [0,6] and [3,3].
a Find the normalized weights and output.

b Now allow only keys 1 and 3. Find the new weights and output.

c Why is zeroing the second old weight without renormalizing wrong?

P27-06-E3. Test information access

A four-position self-attention calculation uses fixed projection matrices and a lower-triangular causal mask.
a Which output rows can change if only the embedding at position 4 changes?

b Does the same guarantee hold with all keys allowed?

c Specify a meaningful automated test and its limit.

P27-06-E4. Audit two automated claims

A hypothetical explanation says: A, doubling all value coordinates doubles the output at fixed Q, K and mask; B, the largest weight proves which biological process caused an observation.
a Assess claim A algebraically.

b Assess claim B from the supplied demonstration.

c Give one code check for A and one evidence requirement for investigating B.

References

Original Python Lecture 3. PDF pages 4–7, matrix prerequisites.

Vaswani and colleagues, Attention Is All You Need. Sections 3.2.1–3.2.3, equation 1; sections 3.4–3.5 for architecture context.
https://arxiv.org/html/1706.03762v7

Python source materials: Phillip Kerger (2023). Course adaptation and additional exposition: Zan Ahmad.