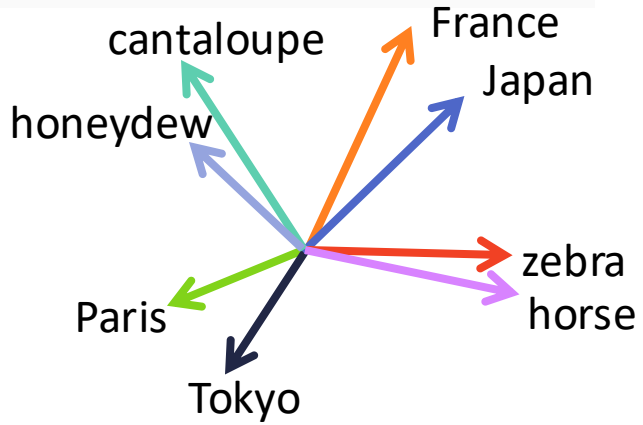
Neural computation models II

Daniel Hsu

COMS 6998-7 Spring 2025

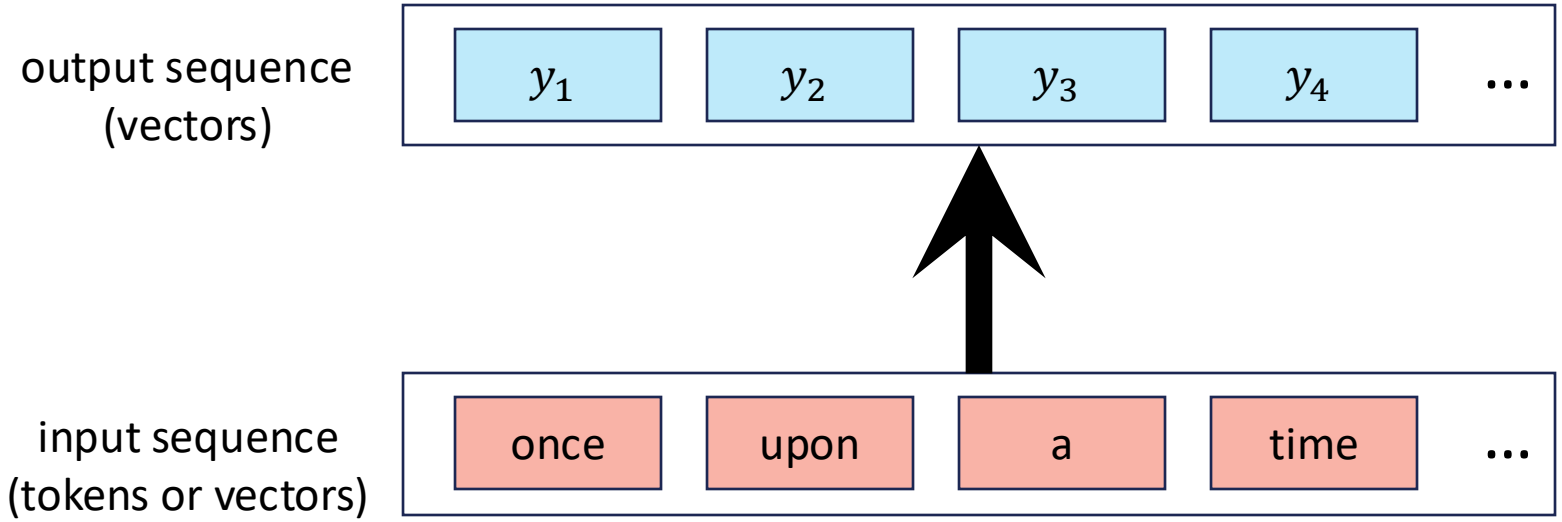
Transformers [Vaswani et al, 2017]



Transformer: a kind of sequence-to-sequence map, formed by compositions of self-attention heads

Ingredients:

1. Ways to embed tokens into vector space
2. Way to for embedded tokens to "interact" and produce new vectors

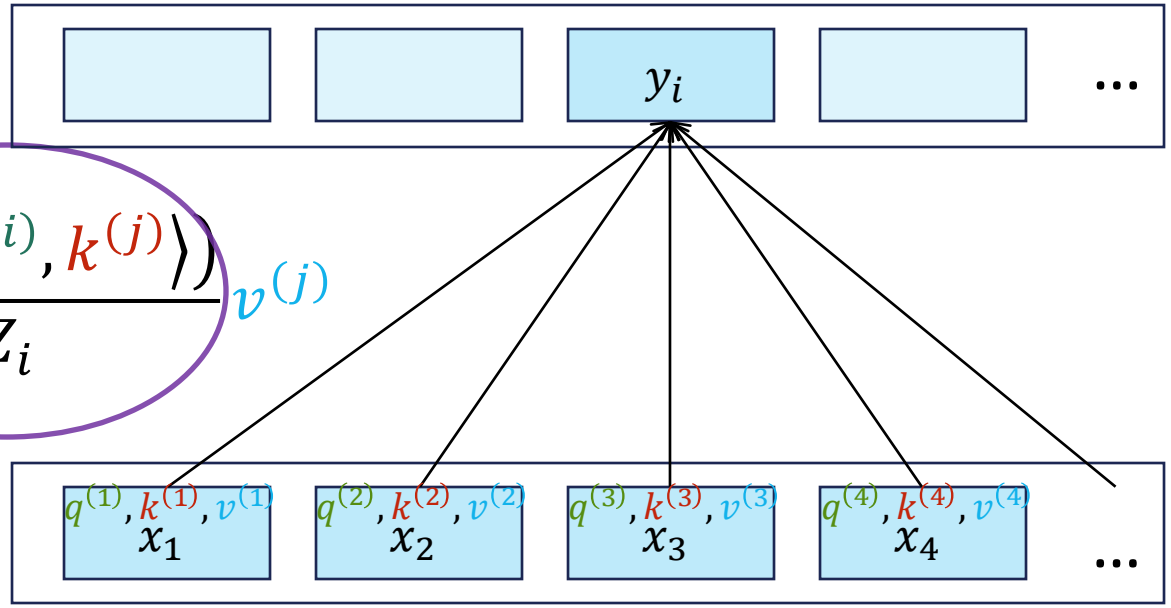


Self-attention head

Token embeddings created using "trained" multilayer Perceptrons (MLPs)

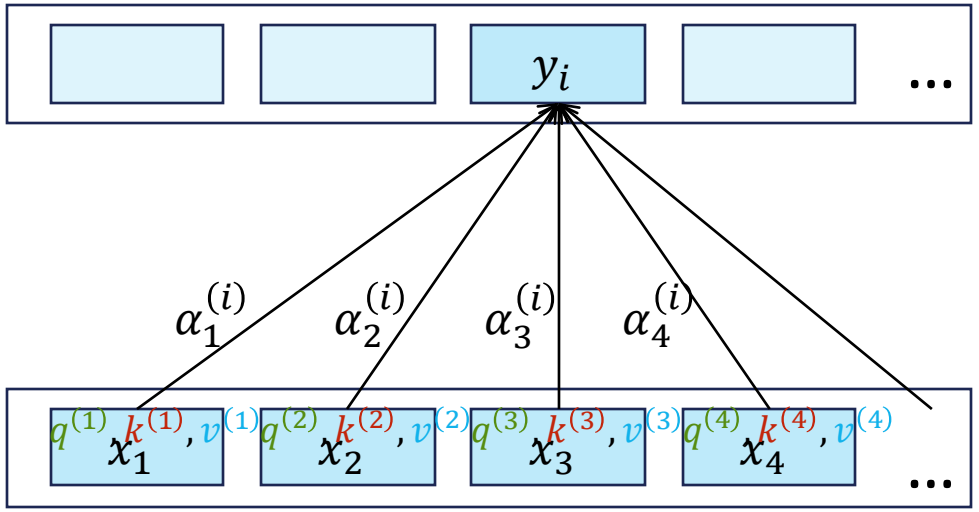
1. Independently create N query/key/value vectors from $x_1, \ldots, x_N$
2. For each $i \in [N]$: i^{th} output y_i = weighted average of all N values, where weights = "softmax" of $\langle i^{\text{th}}$ query, j^{th} key$\rangle$ for all $j \in [N]$

$$y_i = \sum_j \frac{\exp\left(\left\langle q^{(i)}, k^{(j)} \right\rangle\right)}{Z_i} v^{(j)}$$



Outputs $y_1, \ldots, y_N$ can be produced in parallel

Comparison to feedforward neural networks



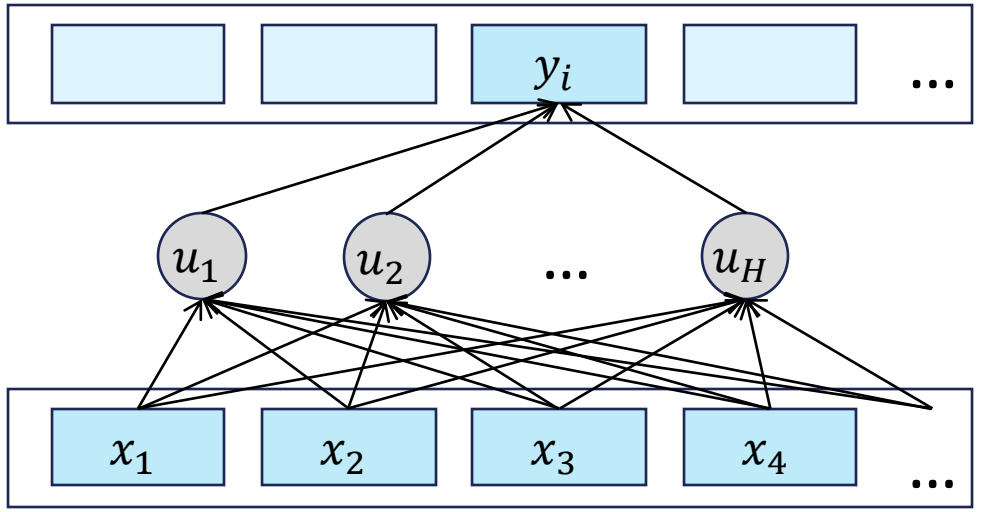
Self-attention head

Shared parameterized mapping

$$x_i \mapsto \left(q^{(i)}, k^{(i)}, v^{(i)}\right)$$

Weights $\alpha_j^{(i)}$ determined via softmax

Universal approximation if embedding dimension $D \to \infty$



Feedforward neural network

Each "weight" is a separate parameter

$$y_i = \sum_{j=1}^{H} A_{i,j}\sigma\left(\sum_{k=1}^{N} W_{j,k}x_k\right)$$

Universal Approximation Bounds for Superpositions of a Sigmoidal Function

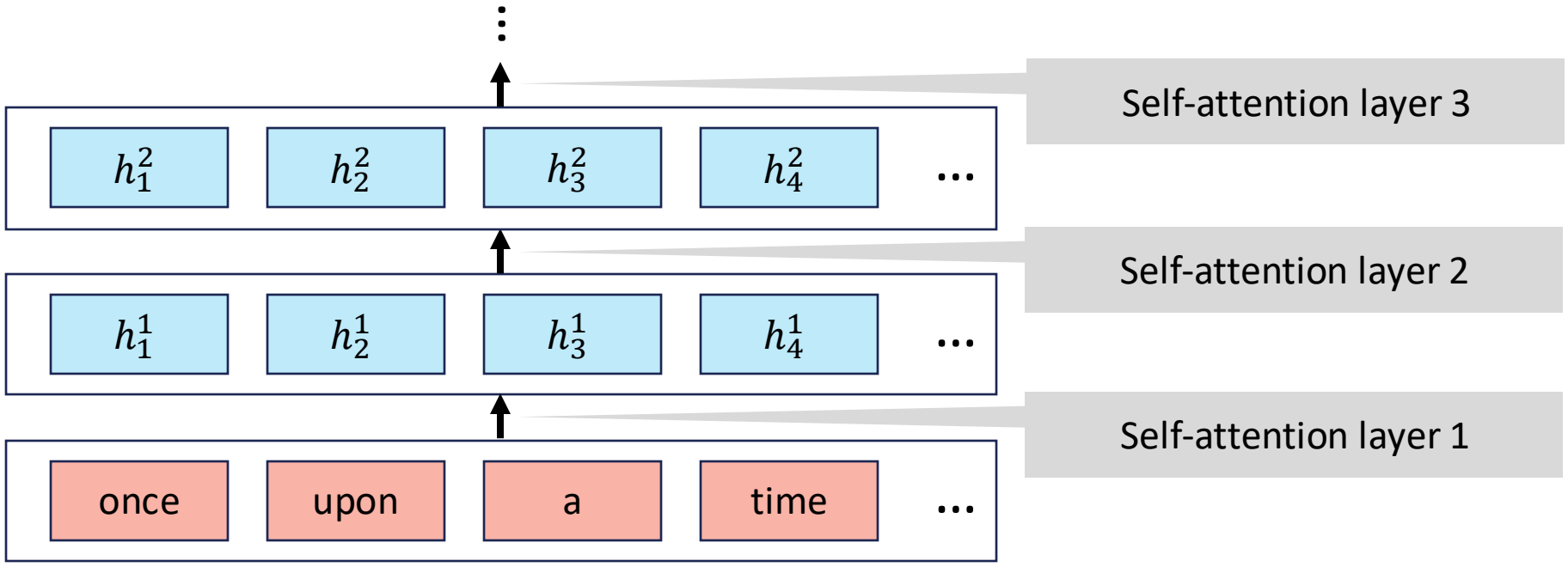
Andrew R. Barron, *Member, IEEE*

(if width $H \to \infty$)

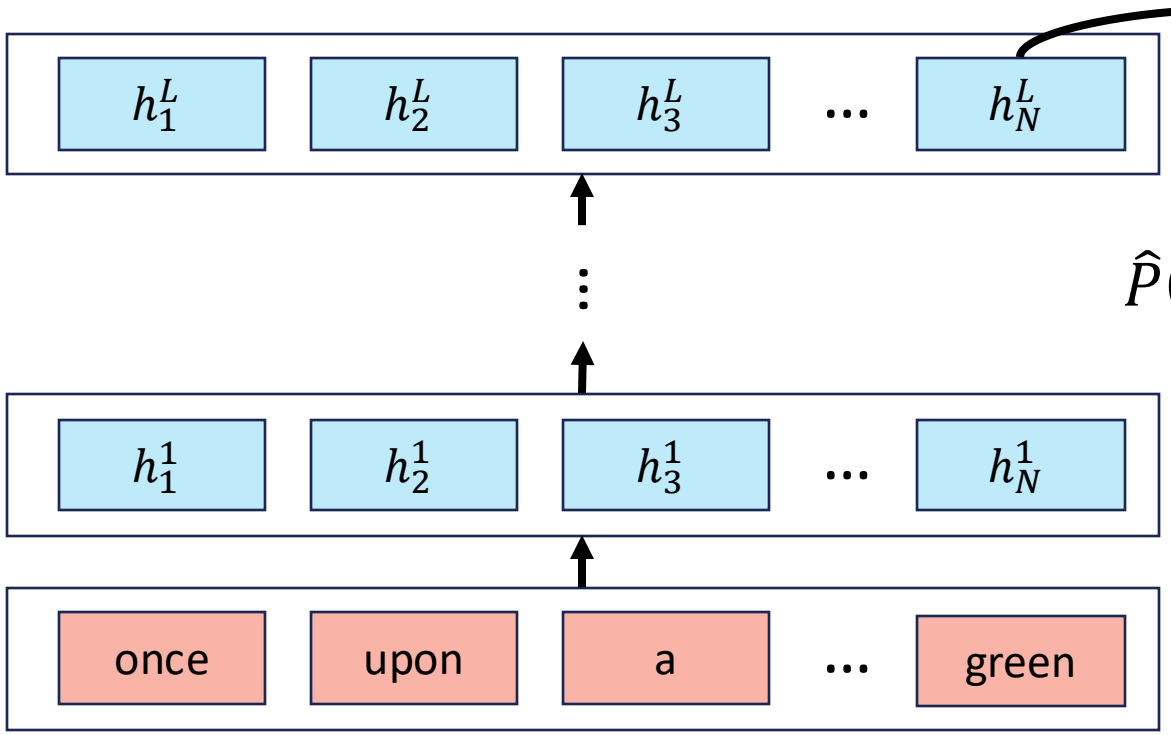
Transformers as compositions

Transformers: compositions of self-attention layers

(layer = one self-attention head, or sum of several self-attention heads)



Use for predicting the next word



$$U(x_{1:N})$$

$$\hat{P}(x_{N+1}|x_{1:N}) \propto \exp\left(T(x_{N+1}) \cdot U(x_{1:N})\right)$$

Other bells-and-whistles

- Self-attention is permutation-equivariant
  - To break permutation-equivariance, typically use positional embeddings
    - Embedding of i^{th} input token x_i may also depend on the position i
- Masked self-attention: To determine "weights" for i^{th} output, only consider subset $S_i \subseteq [N]$ of input tokens (the "unmasked" tokens)
  - Causally-masked self-attention:

$$S_i = \{1,2, \dots, i\}$$

- Skip-connection: Input to next layer is
  - output of current layer …
  - … plus input to current layer

Questions

- What, if anything, is special about the function form of self-attention?
- Two ways to break permutation-equivariance: position embeddings and (causal) masking. Are they interchangeable?
- What is the role of the skip-connection?