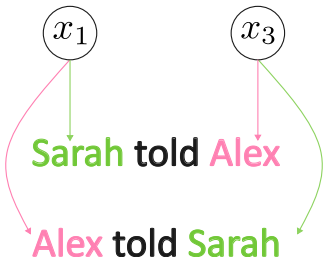
Transformers (Part 2)

Dr. Alireza Aghamohammadi

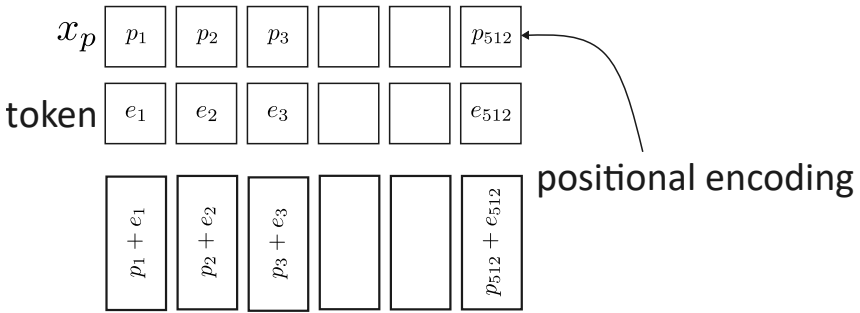
The Importance of Position

- Transformers process each element of the input in parallel.
- The order of elements in the input $(x_1, x_2, \cdots, x_m)$ is crucial.
- Changing the order of elements, such as swapping x_1 and x_3, can significantly alter the meaning of a sentence.
- While the embedding layer transforms each token into a vector, it does not include positional information.
- We should somehow convey the position of each token to the transformer to maintain the correct sequence and meaning.



Positional Encoding Idea

- Each unique position x_p should have a unique vector value.
- We should be able to create position vectors for arbitrary positions, even if we haven't seen sentences longer than 100 tokens during the training process.
- The same token in different positions should have different final representations.
- The values of each position vector should be within a limited range (e.g., between -1 and 1).



Positional Encoding

- Transformers use sine and cosine functions of different frequencies to encode positional information.
- Transformers set the length of the vector for both positional encoding (d) and input embedding to 512.