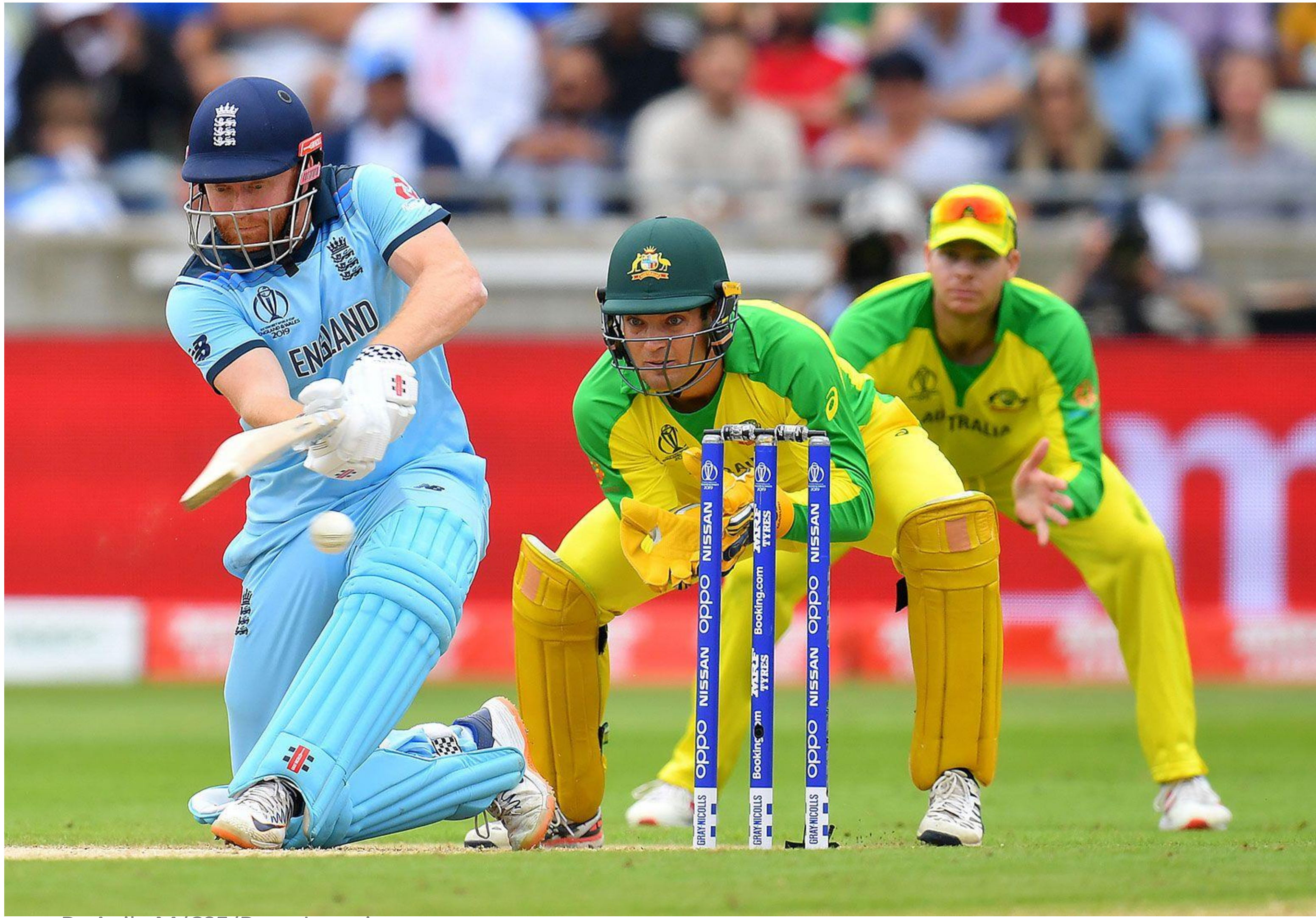
Attention Mechanism, Encoder-Decoder model, Attention over Images, Introduction to Transformers

Unit 4-part2/Deep Learning/VII sem CSE/2024-25

Attention Mechanism

- Example: identify what is the color of the cap wore by player who is batting in the picture?
- (here you focus on imp part of the picture instead of whole picture…attention)

- attention mechanism is just a way of focusing on only a smaller part of the complete input while ignoring the rest.
- In an attempt to borrow inspiration from how a human mind works, researchers in Deep Learning have tried replicating this behavior using what is known as the ‘attention mechanism’.



Attention Mechanism

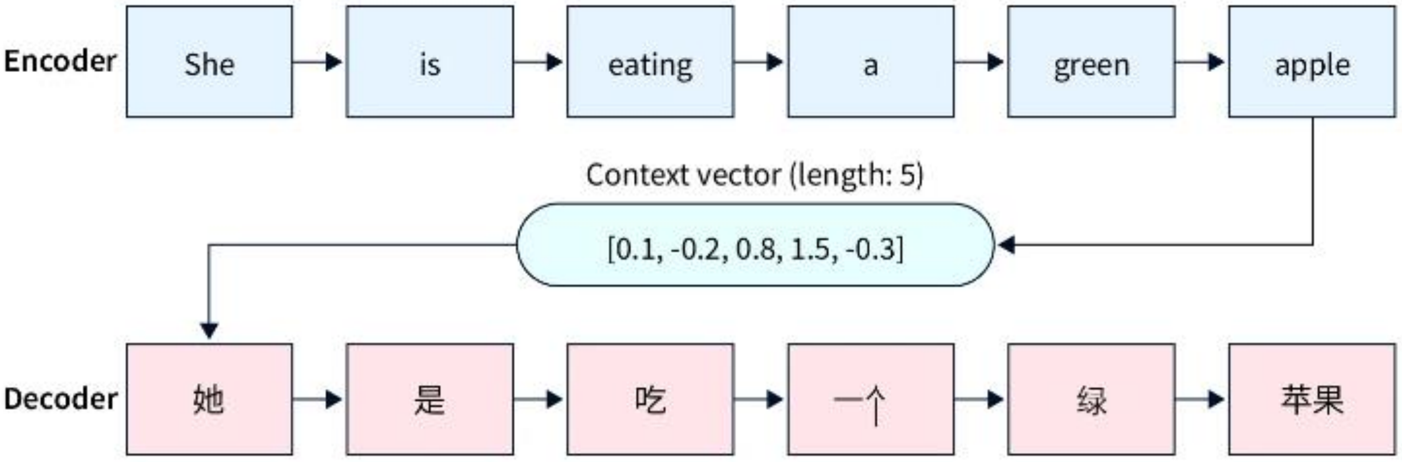
- Attention can be simply represented as a <u>3 step mechanism</u>:

1. Create a probability distribution that rates the importance of various input elements. These input representations can be words, pixels, vectors etc. Creating these probability distributions is actually a learnable task.
2. Scale the original input using this probability distribution such that values that deserve more attention gets enhanced while others get diluted. Kinda like blurring everything else that doesn’t need attention.
3. Now use these newly scaled inputs and do further processing to get focused outputs/results.

Attention Mechanism: Encoder-Decoder Model

- Consider machine translation as an example, where a traditional seq2seq model would be used. Seq2seq models are typically composed of two main components: an **encoder** and a **decoder**.

- The encoder processes the input sequence and represents it as a fixed-length vector (context vector), which is then passed to the decoder.
- The decoder uses this fixed-length context vector to generate the output sequence.
- The encoder and decoder networks are recurrent neural networks like GRUs and LSTMs.



Note: One disadvantage of this approach is the model's inability to remember long sequences because of the **fixed-length context vector**.

Encoder-Decor Model

- To be able to understand the attention mechanism in detail, it is required that you understand Sequence to Sequence models like LSTMs and GRUs.

- The first paper which brought the idea of attention mechanism to the world was Bahdanau et al., 2015. It proposes the encoder-decoder model with an additive attention mechanism.

Types of Attention

1. Self-Attention

- Imagine the sentence below being used as **input** for a machine translation model:

“The driver could not drive the car fast enough because it had a problem.”

- In the above sentence, does the it refer to the driver or the car? Who had the problem?
- For humans, it is a straightforward answer, but for machines, it might not be very clear if they cannot learn the context.
- When doing machine translation, for example, it is important to have attention scores for the source and target sequences, and to have it between the source sequence themselves, thus self-attention.

Types of attention

2. Dot-product Attention

- **Dot-product** Attention computes the attention weights as the dot product of the query and key vectors

$$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$$

3. Scaled dot-product Attention

- **Scaled dot-product** Attention, a variant of dot-product attention that scales the dot product by the square root of the key dimension.

$$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$$

Types of attention

4. Multi-head Attention

- **Multi-head** Attention splits the query, key, and value vectors into multiple heads and applies dot-product attention to each head independently.

5. Self-attention

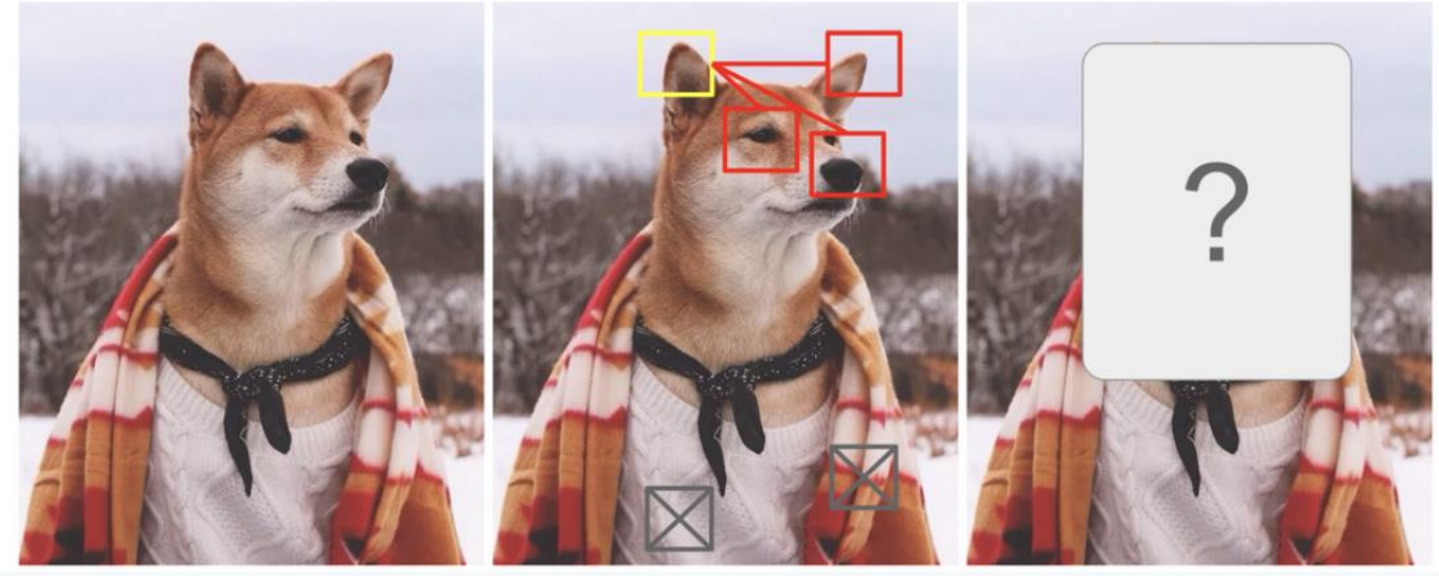
- **Self-attention** is a type of attention mechanism discussed in the previous sections. Here the input sequence is used as both the query and the key.

6. Structured Attention

- **Structured** Attention allows the attention weights to be learned using a structured prediction model, such as a conditional random field.

Attention over Images

- In deep learning, attention can be interpreted as a vector of importance weights.
- When we predict an element, which could be a pixel in an image or a word in a sentence, we use the attention vector to infer how much is it related to the other elements.



If we focus at the features of the dog in the red boxes, like his nose, right pointy ear and mystery eyes, we'll be able to guess what should come in the yellow box.

However, by just looking at the pixels in the gray boxes, you won't be able to predict what should come in the yellow box.

The attention mechanism weighs the pixel in the correct boxes more w.r.t the pixel in the yellow box. While the pixel in the gray boxes would be weighed less.

Attention over Images

- When training an image model, we want the model to be able to focus on important parts of the image. One way of accomplishing this is through **trainable attention** mechanisms (but you already know this, right? Read on..)
- In our case, we are dealing with lesion images, and it becomes all the more necessary to be able to **interpret** the model. It is important to understand which part of the image contributes more towards the cancer being classified benign/malignant.

Intro to Transformers

- A transformer model is a neural network that learns the context of sequential data and generates new data out of it.
- To put it simply:

A transformer is a type of artificial intelligence model that learns to understand and generate human-like text by analyzing patterns in large amounts of text data.

- Transformers are a current state-of-the-art NLP model and are considered the evolution of the encoder-decoder architecture.
- However, while the encoder-decoder architecture relies mainly on Recurrent Neural Networks (RNNs) to extract sequential information, Transformers completely lack this recurrency.

Transformers

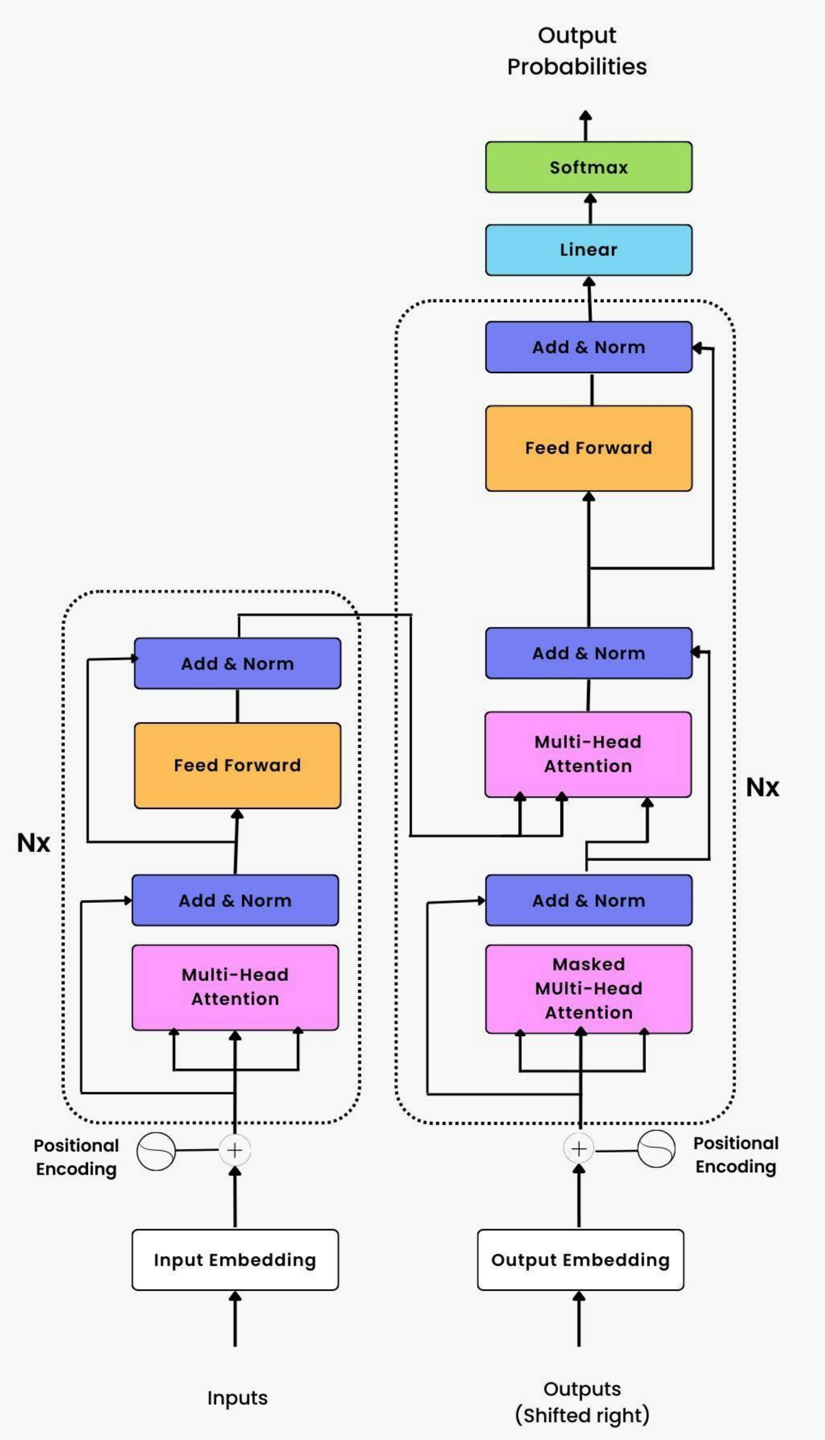
- Transformers came into action in a 2017 Google paper as one of the most advanced models ever developed. This has resulted in a wave of advances called "Transformer AI" in machine learning.
- In a paper published on August 2021, researchers from Stanford identified Transformers as "foundation models" because they believe it will transform artificial intelligence.

shift from RNN models like LSTM to Transformers for NLP problems

- Transformers' ability to assess both of them by taking advantage of the Attention mechanism improvements:

1. *Pay attention to specific words, no matter how distant they are.*
2. *Boost the performance speed.*

- Transformers were inspired by the encoder-decoder architecture found in RNNs. However, Instead of using recurrence, the Transformer model is completely based on the Attention mechanism.
- Besides improving RNN performance, Transformers have provided a new architecture to solve many other tasks, such as text summarization, image captioning, and speech recognition.



Transformers

This figure illustrates the Transformer deep learning model's overall architecture.
There are two main components of the Transformer:

•The encoder stacks — Nx identical encoder layers (in the original published paper, Nx = 6).

•The decoder stacks — Nx identical decoders layers (in the original published paper, Nx =6)

Models do not include recurrences or convolutions, so there is an extra layer of positional encoding between the encoder and decoder stacks to take advantage of the order of the sequence.