Attention mechanism enable neural network to selecting focus on specific part of input sequence.

Transformer

↳ type of neural network.

↳ To handle sequence to sequence data.

ex → machine translation
question answering
text summary.

word embedding → word embedding is a natural language processing technique that converts word into numerical vectors, allowing computers to understand their meaning and relationships.

these vectors are dense.
↳ means they contain many non-zeros value.

ex → TF-IDF, word2vec, and Glove.

static word embedding → static word embedding provide a single, fixed vector for each word.

static embedding are simple and faster

contextual word embedding →

contextual word embeddings generate a unique vector for a word based on the specific sentence it appears in.

contextual embeddings are more accurate because they can capture different meanings of a word, such as the word "bank" in "river bank" versus "money bank"

Self attention mechanism

↳ self attention is a way to convert static embedding to dynamic contextual embedding.

positional encoding provides transformers with information about the order of tokens in a sequence.

multi-head attention is an extension of the self attention mechanism, used in the transformer architecture, which allows the model to focus on different parts of the input sequence from multiple perspectives.

~~Layer, Normalization is a technique used~~ to ~~normalize the inputs across the features for each data~~ point.

Layer normalization in transformer is a crucial technique that stabilizes and speed up model training by normalizing the activations of each layer to have a mean of zero and a standard deviation of one.

Feed forward layers→

↳ This layer applied in both part of the transformer, encoder and decoder.

This layer represent non-linearity and capture complex pattern of the network.

Softmax layer →

The softmax layer in transformer conver a vector of raw scores into a probability distribution.

There are two roles of softmax lay

1. In the attention mechanism
2. final output classification

=) within the self_attention layer, it calculates a probability distribution over the input tokens.

=) In the final layer, it converts the model's final output into a probability for each possible class in a multi-class classification task.

masked multi-head attention is used in the decoder of a transformer to prevent the model from seeing the future when predicting the next token in a sequence.

1 to n, n+1

What are diffusion models→

Diffusion models are a class of generative AI models that generate high resolution images of varying quality.

They work by gradually adding ~~noise~~ gaussian noise to the original data in the forward diffusion process and then learning to remove the noise in the reverse diffusion process.

They are latent variable model

① Forward diffusion process → The forward diffusion process is the markov chain of diffusion steps in which we slowly and randomly add noise to the original data

② Reverse diffusion process → the reverse diffusion process tries to reverse the diffusion process to generate original data from the noise.

$$q(x_t|x_{t-1}) = N(\underset{\text{output Image (noisy Image)}}{x_t} ; \underbrace{\sqrt{1-\beta_t}\cdot x_{t-1}}_{\text{mean}}, \underbrace{\beta_t I}_{\text{Variance}}$$

Normal distribution / gaussian distribution

$$N(0, 1)$$

$\beta_t \rightarrow$ control the amount of noise add in each time step

$$\underset{\text{less Noisy Image}}{P_\theta}(\underset{\text{Noisy Image}}{x_{t-1}}|x_t) = N(\underset{\text{output}}{x_{t-1}} ; \underbrace{\mu_\theta(x_t, t)}_{\text{mean}}, \underbrace{\Sigma_\theta(x_t, t)}_{\text{Variance}})$$

stable deffusion works in a compressed latent space to be significantly faster and require less computational power, while maintaining high quality outputs.

Tokenization

Tokenization is the process of breaking down text into smaller unit called tokens.

These tokens can be words, subwords, or even characters, depending on the approach.

Tokenization is fundamendal in Natural language processing (NLP) and generative AI because it converts raw text into a form that model can process and understand.

Text → Generative AI is fascinating

tokens = ["Generative', 'AI', 'is', 'fascinating']

subword tokenization →

Subword tokenization involves breaking word into smaller, meaningful subword unit.

word → fascination

Subword tokens: ['fas', 'cina', 'tion']

cosine similarity

cosine similarity is a metric used to measure how similar two vectors are, regardless of their magnitude

it is used in text analysis, recommend system and clustering task

$$\text{cosine similarity} = \frac{A \cdot B}{|A|\,|B|}$$