Le traitement du langage naturel par transformers illustré par un exemple pour la classification de texte

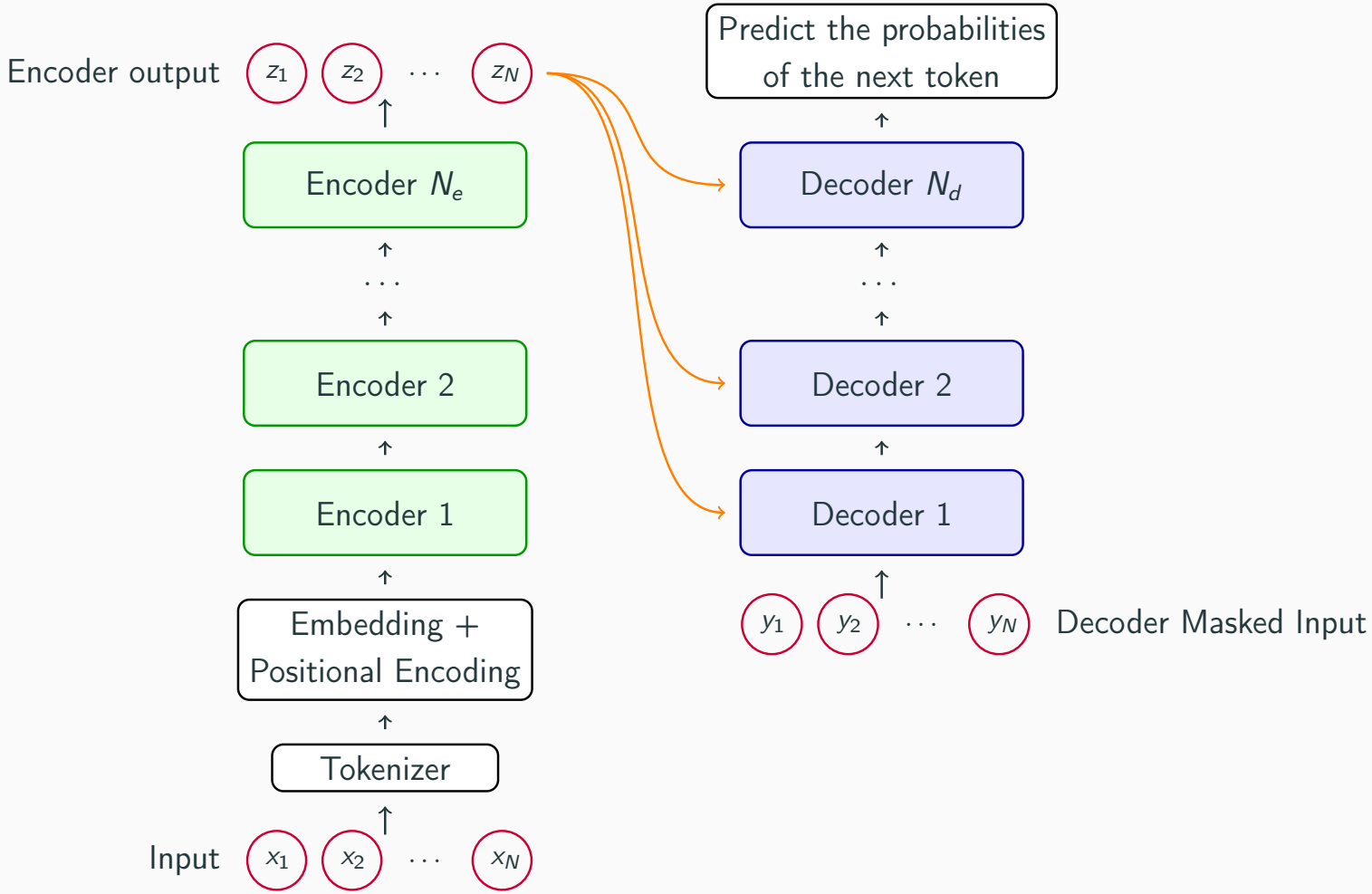
Cerisara Nathan, MPI

23 mai 2023

Sommaire

L'architecture Transformer

Schéma de l'architecture dans le cas de la génération :



Vectorisation du texte : Tokenisation du texte

La tokenisation est le processus de découpage d'une séquence de texte en unités discrètes appelées "tokens". Dans BERT, la tokenisation est réalisée à l'aide du tokenizer BERT inclus dans la bibliothèque Transformers. Le tokenizer prend en entrée une séquence de texte et renvoie une liste de tokens correspondants.

Exemple : la phrase "C'est une phrase d'exemple." peut être tokenisée en ["C", "'", "est", "une", "phrase", "d", "'", "exemple", "."].

Vectorisation du texte : Embeddings

Après la tokenisation, chaque token est converti en un vecteur de nombres réels appelé "embedding". Les embeddings captent les informations sémantiques et syntaxiques des tokens dans un espace vectoriel continu. Les embeddings de BERT sont appris lors de la phase de pré-entraînement du modèle sur de grandes quantités de données textuelles.

Exemple

Phrase d'entrée : "C'est une phrase d'exemple."

Tokens : ["[CLS]", "C", "'", "est", "une", "phrase", "d", "'", "exemple", ".", "[SEP]"]

Embeddings :

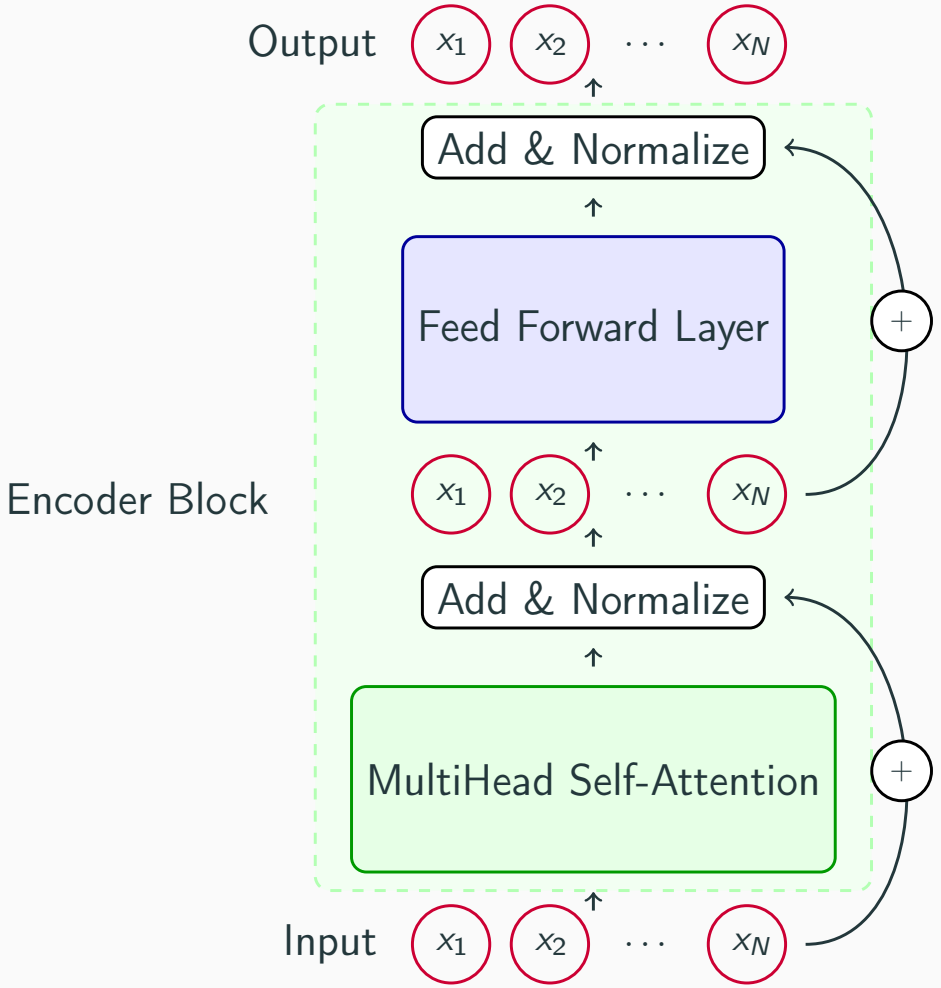
$[embedding_C, embedding_{apostrophe}, embedding_{st}, embedding_{ne}, e$

Vectorisation du texte : Positional Encoding

Positional Encoding : Adding positional information to token embeddings. Formula :
$\text{Encoding}(pos, 2i) = sin(pos/(10000^(2i/d_model)))$,
$\text{Encoding}(pos, 2i+1) = cos(pos/(10000^(2i/d_model)))$.
Example : Calculating positional encoding for each token and embedding dimension.

La partie Encodeur

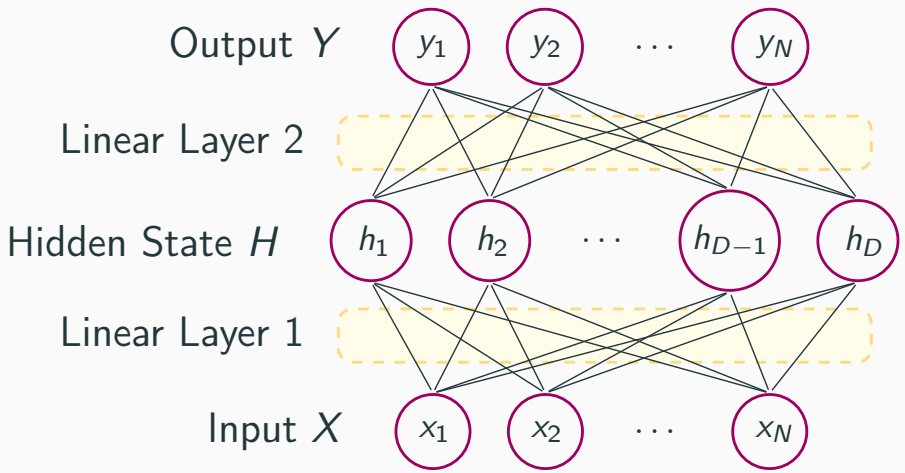
Schéma d'un block de la partie Encodeur de l'architecture Transformer :



Matrice d'attention

Attention Matrices : Capture relationships and dependencies between tokens. Calculation : Applying matrix operation between query, key, and value embeddings. Interpretation : High values indicate strong correlations between tokens.

Le réseau Feed Forward



Couche linéaire :

$$X \mapsto X \cdot W^\top + B$$

Application Personnelle : Objectifs & Rapport à la ville

Le modèle BERT

La structure du réseau de neurone utilisée

Les données et l'apprentissage

Les résultats

Annexes

S