TRABALHO – TRANSFORMER PARA ÁREA DE LINGUAGEM NATURAL

CONTRIBUIÇÕES

Objetivo:

- Treinar algoritmos de aprendizado de máquina utilizando técnicas de otimização de rede neural artificial e comparar a performance dos modelos para classificação.

REFERÊNCIAS

Referências

- Ashish, V., et all, Attention Is All You Need
- Gowtham, R., et all, Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages
- Pimentel, C.H.M.; Pires, T.B., Treinamento e análise de um modelo de tradução automática baseado em Transformer
- CodeEmporium (https://www.youtube.com/watch?v=QCJQG4DuHT0&list=PLTl9hO2Oobd97qfWC40gOSU8C0iu0m2l4)

CONCEITOS

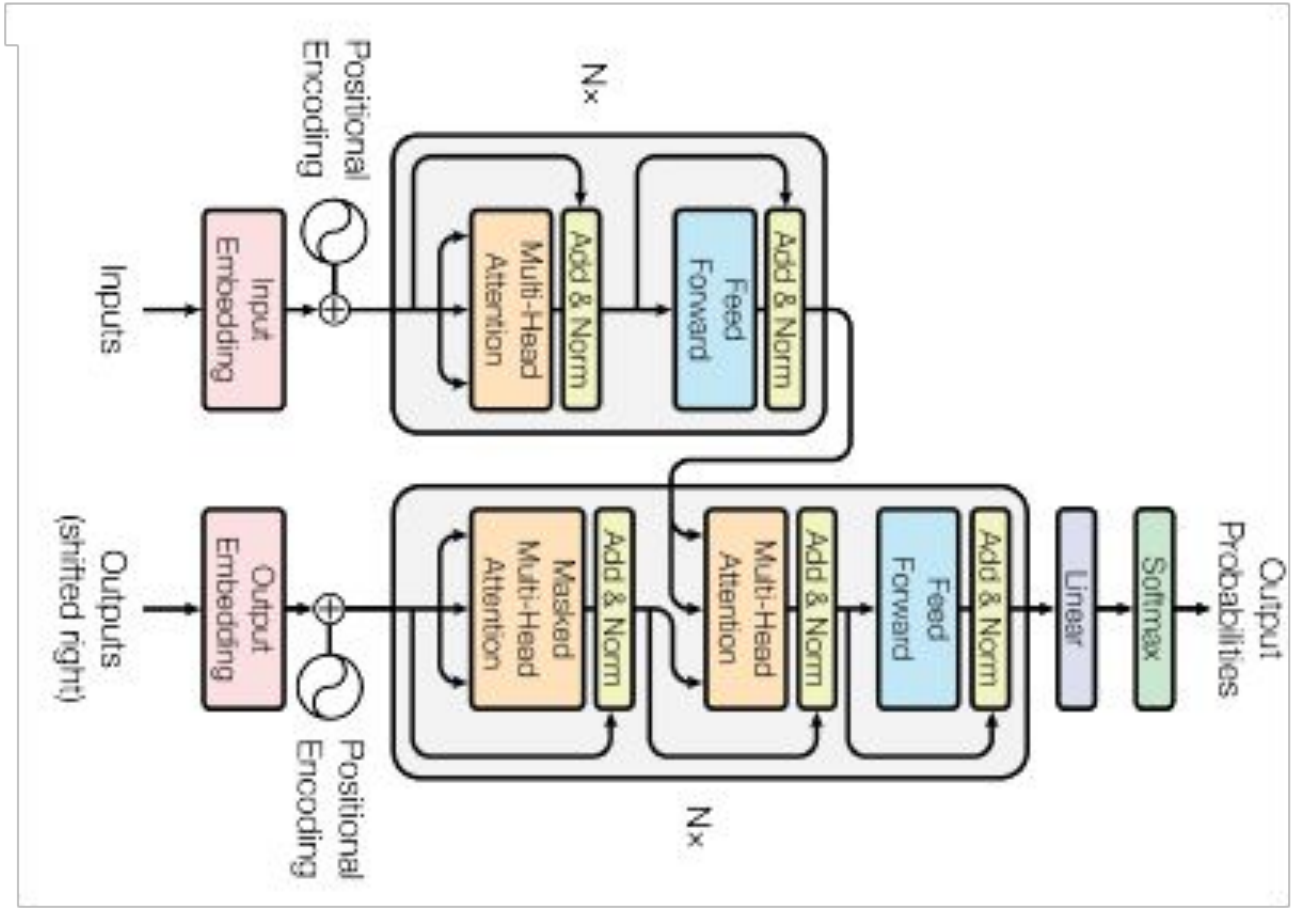

Figure 1: The Transformer - model architecture.

CONCEITOS

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

CONCEITOS

kannada_vocabulary = [START_TOKEN, ' ', '!', '"', '#', '$', '%', '&', "'", '(', ')’
, '*', '+', ',', '-', '.', '/', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ':’
, '<', '=', '>', '?', '@', 'ಂ', 'ಅ', 'ಆ', 'ಇ', 'ಈ', 'ಉ', 'ಊ', 'ಋ', 'ಎ', 'ಏ', 'ಐ', 'ಒ', 'ಓ', 'ಔ', 'ಕ', 'ಖ', 'ಗ', 'ಘ', 'ಙ', 'ಚ', 'ಛ', 'ಜ', 'ಝ', 'ಞ', 'ಟ', 'ಠ', 'ಡ', 'ಢ', 'ಣ', 'ತ', 'ಥ', 'ದ', 'ಧ', 'ನ', 'ಪ', 'ಫ', 'ಬ', 'ಭ', 'ಮ', 'ಯ', 'ರ', 'ಱ', 'ಲ', 'ಳ', 'ವ', 'ಶ', 'ಷ', 'ಸ', 'ಹ', '಼', 'ಽ', 'ಾ', 'ಿ', 'ೀ', 'ು', 'ೂ', 'ೃ', 'ೄ', '೅', 'ೆ', 'ೇ', 'ೈ', '೉', 'ೊ', 'ೋ', 'ೌ', '್', 'ೕ', 'ೖ', 'ೞ', 'ೠ', 'ೡ', 'ೢ', 'ೣ', '೦', '೧', '೨', '೩', '೪', '೫', '೬', '೭', '೮', '೯', 'ೱ', 'ೲ', PADDING_TOKEN, END_TOKEN]

```
english_vocabulary = [START_TOKEN, ' ', '!', '"', '#', '$', '%', '&', "'", '(', ')’,
'*', '+', ',', '-', '.', '/', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9’,
':', '<', '=', '>', '?', '@’, '[', '\\', ']', '^', '_', '`', 'a', 'b', 'c', 'd', 'e’,
'f', 'g', 'h', 'i', 'j', 'k', 'l’, 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v’,
'w', 'x’, 'y', 'z’, '{', '|', '}', '~', PADDING_TOKEN, END_TOKEN]
```

MATERIAIS E MÉTODOS

- Dataset com 300000 linhas de textos em inglês e Kannada para ser usado para treinamento e validação
- Uso da arquitetura transformer com dropout
- Inicialização do transformer:

  d_model = 512
  batch_size = 30
  ffn_hidden = 2048
  num_heads = 8
  drop_prob = 0.1
  num_layers = 1
  max_sequence_length = 200
  kn_vocab_size = len(kannada_vocabulary)

IMPLEMENTAÇÃO

- Códigos: https://drive.google.com/file/d/1oYFz1om72eD3UniGMI5epM_d2iNDCZeY/view?usp=sharing
- A implementação do transformer encontra-se no arquivo transformer.py
- O arquivo Transformer_Trainner_Kannada.ipynb trata do tratamento dos dados, da inicialização do transformer, do treinamento e teste.

IMPLEMENTAÇÃO

- A implementação do transformer encontra-se no arquivo transformer.py
- O arquivo Transformer_Trainner_Kannada.ipynb trata do tratamento dos dados, da inicialização do transformer, do treinamento e teste.

RESULTADO DO ARTIGO

Após a coleta dos textos, foram computados 134.943 *tokens* e 6.453 *types* presentes nos textos em inglês e 143.285 *tokens* e 8.932 *types* em francês. Ao final foram alinhadas 5.494 frases em inglês-francês. A partir disso, o modelo treinado recebeu o *score* sacreBLEU de 7,6467.

Tabela 2. Comparação dos scores sacreBLEU atribuídos às frases geradas pelo modelo treinado e pelo Google Tradutor.

Chave	Score sacreBLEU Modelo Treinado	Score sacreBLEU Google Tradutor
316	32,5	32,6
384	29,5	58,1
635	71,9	50,7
796	22,8	23,2
852	8,3	6,3
950	20,6	12,6
965	21	63,9
1013	82,4	91,2
1166	7,3	7,1
1377	22	26,3
1390	33	7,8
1399	49,2	11,4
1411	49,8	36,1
1418	23,4	10
1437	5,7	14,1
1451	14,4	9,4
1455	29	24,6
1471	37,5	33,5
1486	41,4	34,8
1520	31.9	27,2
Média	31,7	29,0

Fonte: Elaboração própria.

TRABALHOS FUTUROS

Fazer uso do transformer criado para aplicar nas nossas pesquisas.