Transformer and Advanced Models for Transfer Learning in NLP

Akash (SC22M083)
Machine Learning and Computing
Department of Mathematics
Indian Institute of Space Science and Technology

Under The Guidance Of

Dr. S. Sumitra (Associate Professor)
Department of Mathematics
Indian Institute of Space Science and Technology

Ms. Sanjeena Menezes (Senior Manager)
Architecture and Networking
Continental Automotive Components (India) Pvt. Ltd.

December 2023

Contents

Abstract

Embark on a comprehensive exploration of Transformer architecture and advanced models for transfer learning in Natural Language Processing (NLP). Delve into the intricate details of the Transformer architecture, with a special focus on the heart of its attention mechanism. Uncover the essence of transfer learning in NLP and its significance. Journey through advanced models like GPT, BERT, and BART, understanding their architecture, pretraining and fine-tuning approaches. This seminar presentation provides a nuanced understanding of the core components shaping contemporary language processing, offering insights into the transformative landscape of cutting-edge NLP models.

Introduction

- NLU and NLG have been extensive research topics, aiming to enable machines to comprehend and generate language akin to humans.
- Historically challenging, replicating human-like language understanding and generation in machines is now achievable due to advancements in AI techniques.
- Contemporary technology, particularly AI, has significantly progressed, empowering machines to perform tasks such as Question Answering, Language Translation, Story Writing, and Document Summarization.
- While solving specific language tasks like machine translation demands labeled data, trained models are often not versatile for other language-related tasks, unlike in computer vision.
- Drawing inspiration from transfer learning in computer vision, can we build models in NLP that learn from vast data and apply knowledge to diverse language-related tasks?
- With abundant unlabeled textual data, training models based on transformer architecture in an unsupervised manner becomes feasible, enabling them to grasp language understanding.
- Utilizing pretrained models allows for transfer learning in language, where models trained on unsupervised data can be fine-tuned in a supervised manner for downstream NLP tasks.

Transformer

- The Transformer, introduced by Vaswani et al. in the seminal paper Attention is All You Need.
- It is sequence-to-sequence model.
- SOTA Architecture in the field of NLP outperformed existing model for machine translation task.
- Offer parallel processing reducing time during training and inferencing stages.
- Consists two block Encoder and Decoder.
- Each block consist Attention and Feed Forward Network followed by layer addition and layer norm.
- The heart of transformer is attention mechanism.
- Offering significant advantages in capturing long range dependency.

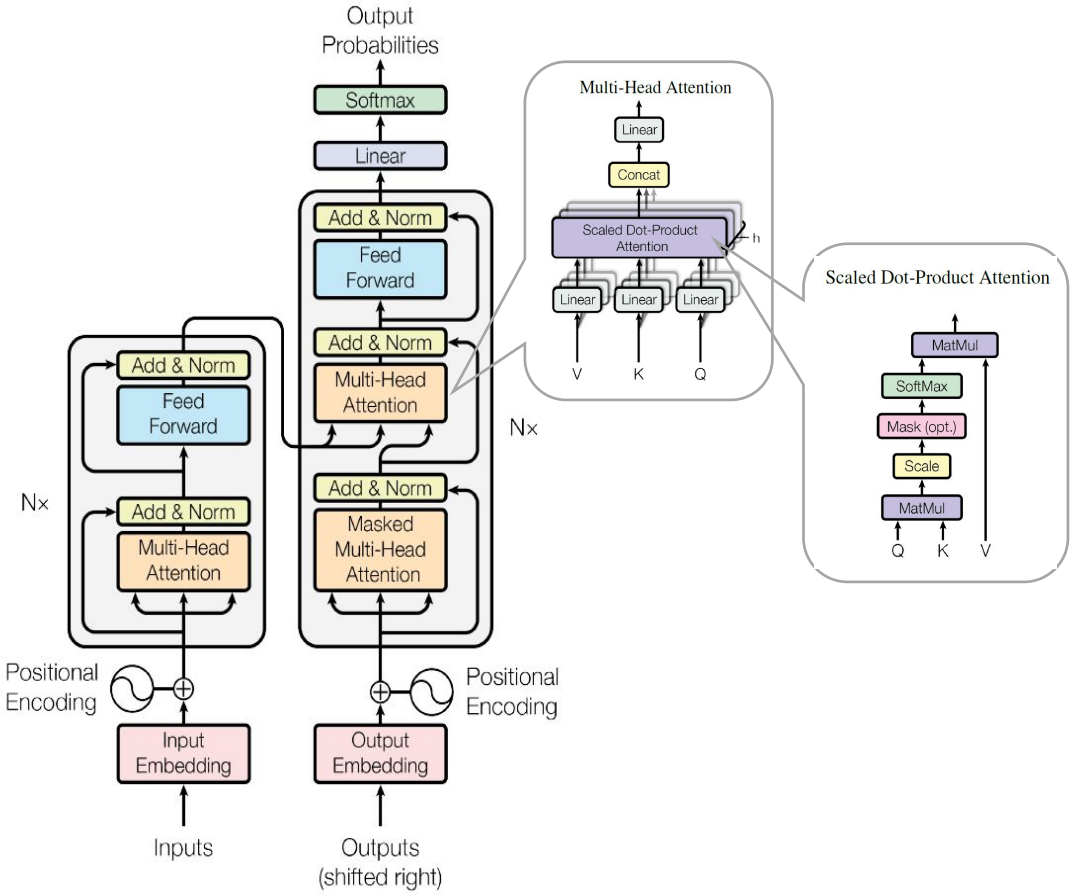


Fig. The Transformer - Model Architecture

Overview and Architecture of Transformer cont..

Key Components

Input Embedding: Transforms input tokens into meaningful vector representations.

Positional Encoding: Injects sequence information into transformer's neural network.

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

Encoder: Encoder in transformer analyzes input sequence, applying self-attention, producing contextualized representations.

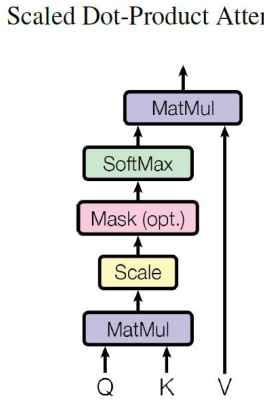
Decoder: The decoder in a transformer generates output sequences by attending to the encoded input.

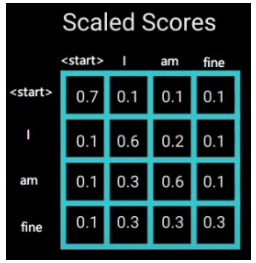
Attention: Attention is a mechanism allows the model to focus on (attend to) the most relevant parts of the input. Attention mechanism improving context understanding.

Overview and Architecture of Transformer cont..

Scaled Dot-Product Attention: It is a mapping (function) which take query, key and value as input and calculate attention score.

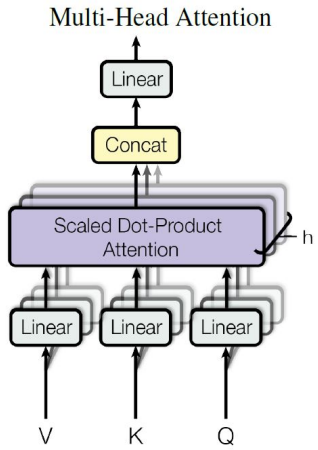
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$





Multi-head Attention: Multi-head Attention combines diverse weighted views, enhancing model's ability to capture complex relationships in data.

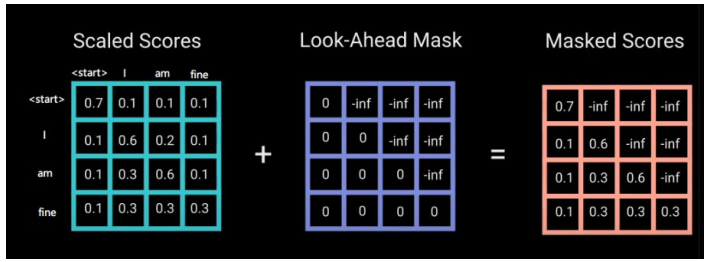
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

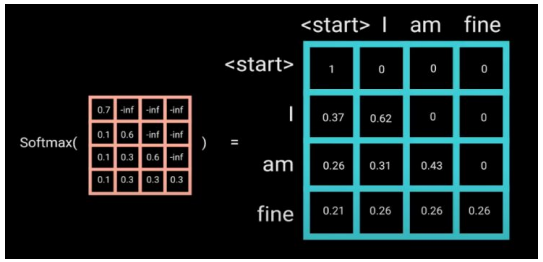


Overview and Architecture of Transformer cont..

Masked Multi-head Attention: In the decoder, masked multi-head attention is used to prevent attending to future tokens. A masking matrix is applied to the softmax input.

$$MaskedAttention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + mask\right)V$$





Add & Layer Norm: Transformer employs "Add and Layer Norm" in each sub-layer, promoting stable training by mitigating vanishing/exploding gradients and improving convergence.

$$Add\&Norm(x) = LayerNorm(x + Sublayer(x))$$

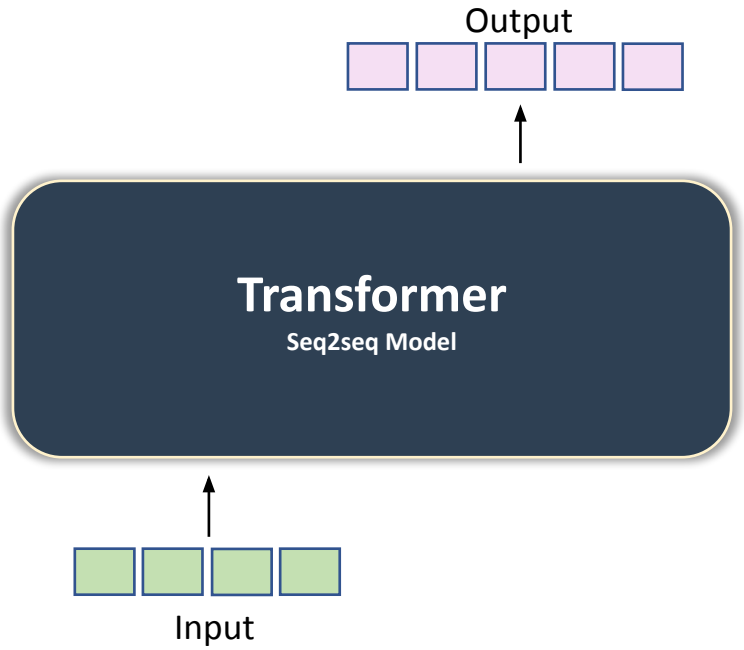
Feed-Forward Network: Transformer's Feedforward Neural Network applies non-linear transformations, capturing intricate patterns in data, enhancing the model's representational capacity.

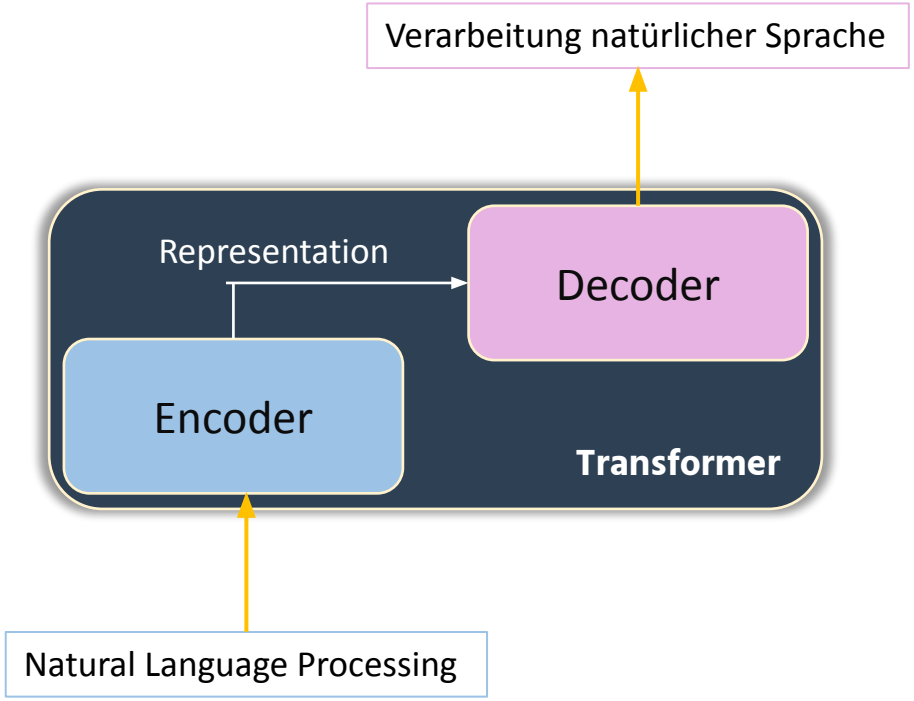
$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Overview and Architecture of Transformer cont..

Linear Layer: Linear layers are used for transformations, typically followed by layer normalization. It transform decoder output vector to dimension of vocabulary size.

Softmax Layer: Transformer's Softmax Layer converts decoder output to probability distribution, crucial for sequence generation tasks, ensuring well-formed and meaningful outputs.



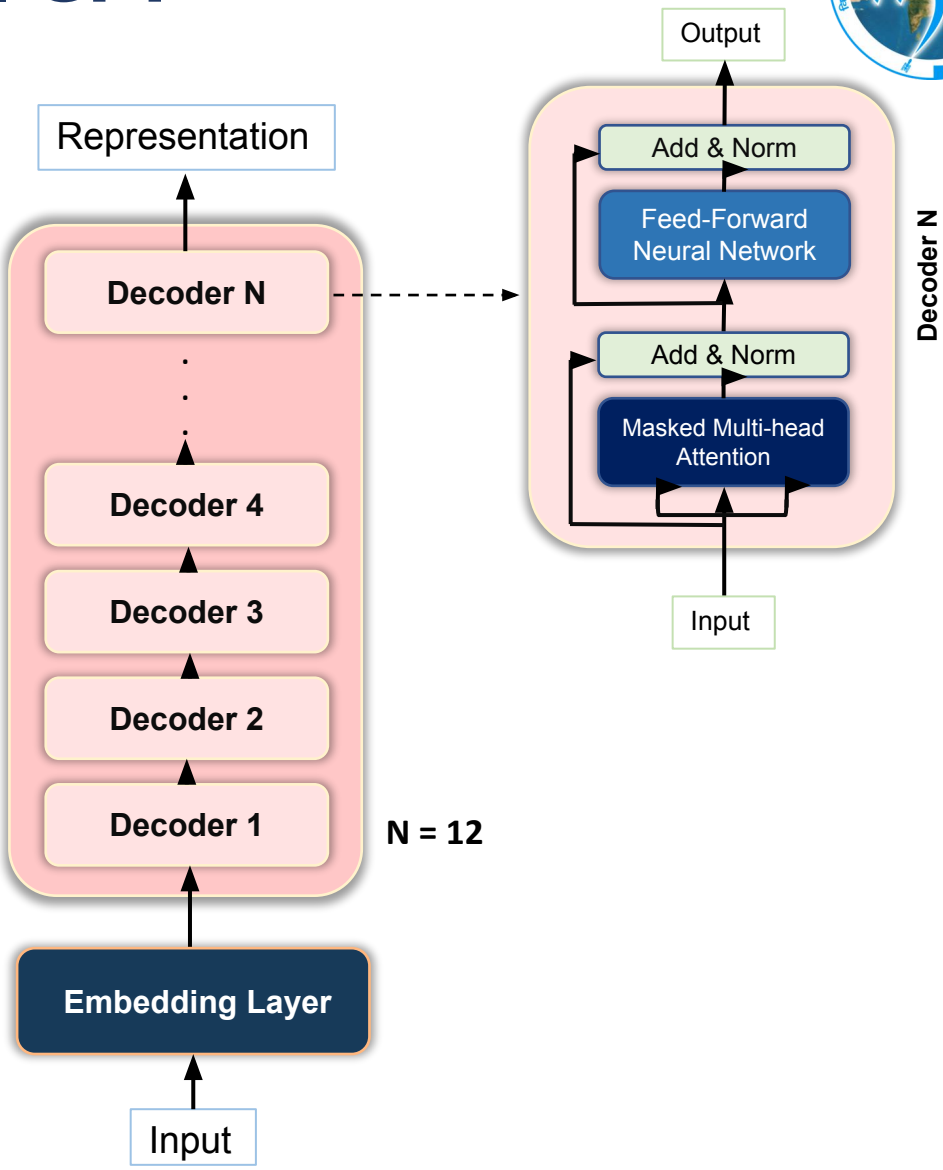


Transfer Learning in NLP

- Transfer learning in NLP is like using a well-trained brain to help with new language tasks.
- Imagine learning from lots of stories before tackling a new writing assignment. In the same way, we use pretrained models to learn from lots of language data before working on specific tasks like understanding sentiments in texts. This helps us make smart language models.
- This approach is changing how we do language research. Shared pretrained models are like sharing big libraries of knowledge, speeding up how we create new language models and use them in different situations.
- It's a bit like how we learn general things before applying them to specific problems, just like we learn general math before solving math problems.
- So, transfer learning in NLP makes language models smarter and more useful by learning from a wealth of language data before focusing on specific tasks.

Overview and Architecture of OpenAI GPT

- GPT Introduced by OpenAI in 2018, revolutionized AI.
- Groundbreaking approach to NLP and NLU.
- Based on two existing idea: Transformer and unsupervised pretraining.
- Decoder only Transformer model.
- 12-layer, 768-hidden, 12-heads, 117M parameters.
- Marked a breakthrough in unsupervised learning.
- Showcased impressive text generation capabilities and versatility.
- Marked a significant milestone in the development of AI, showcasing the potential of pre-trained models.
- Success of GPT laid the foundation to development of more advanced language models.



Overview and Architecture of OpenAI GPT Cont..

GPT Pretraining:

Trained on **BookCorpus Dataset** contains more than 7000 books from various genres.

Unsupervised pre-training

$\mathcal{U} = \{u_1, u_2, \ldots, u_n\} \rightarrow$ unsupervised corpus of tokens

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

(k: size of the context window)

Transformer decoder

$U = (u_{-k}, \ldots, u_{-1}) \rightarrow$ context vector of tokens

$$h_0 = UW_e + W_p$$

(W_e: token embedding matrix, W_p: position embedding matrix)

$$h_l = \texttt{transformer\_block}(h_{l-1}), l = 1, \ldots, L$$

(L: number of layers)

$$P(u) = \texttt{softmax}(h_L W_e^T)$$

BookCorpus Dataset

Overview and Architecture of OpenAI GPT Cont..

GPT Finetuning:

GPT can be finetuned for various NLP Downstream task in a supervised manner.

Supervised fine-tuning

$\mathcal{C} \rightarrow$ labeled dataset

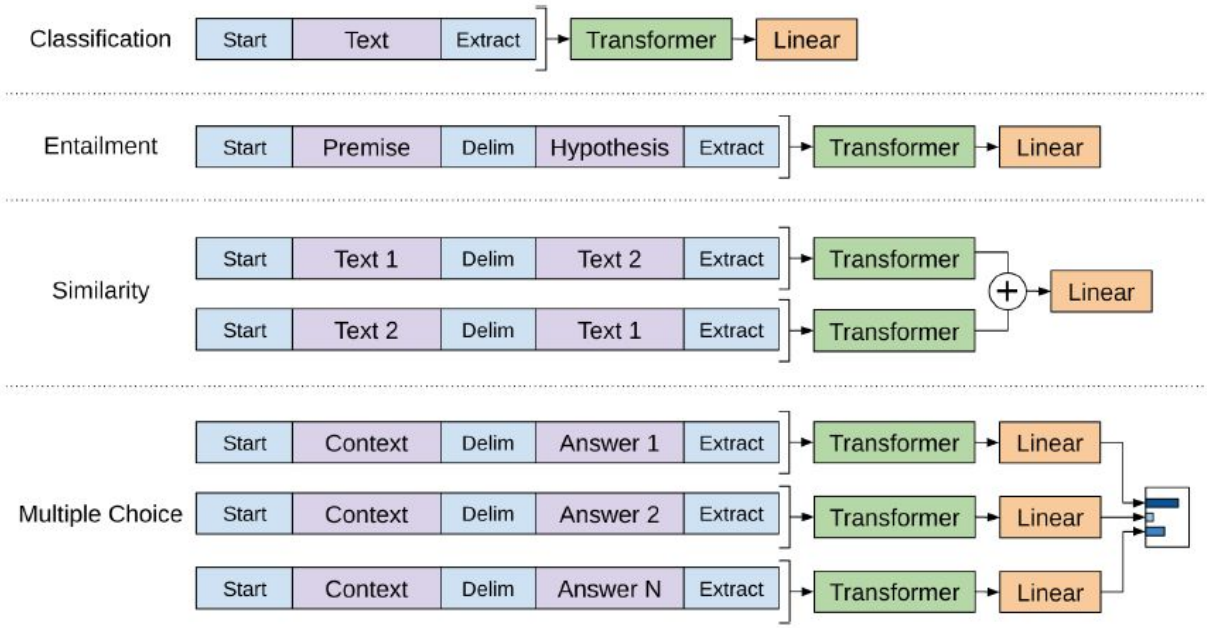
$x^1, \ldots, x^m \rightarrow$ input tokens

$y \rightarrow$ label

$h_L^m \rightarrow$ final transformer block's activation

$$p(y|x^1, \ldots, x^m) = \text{softmax}(h_L^m W_y)$$

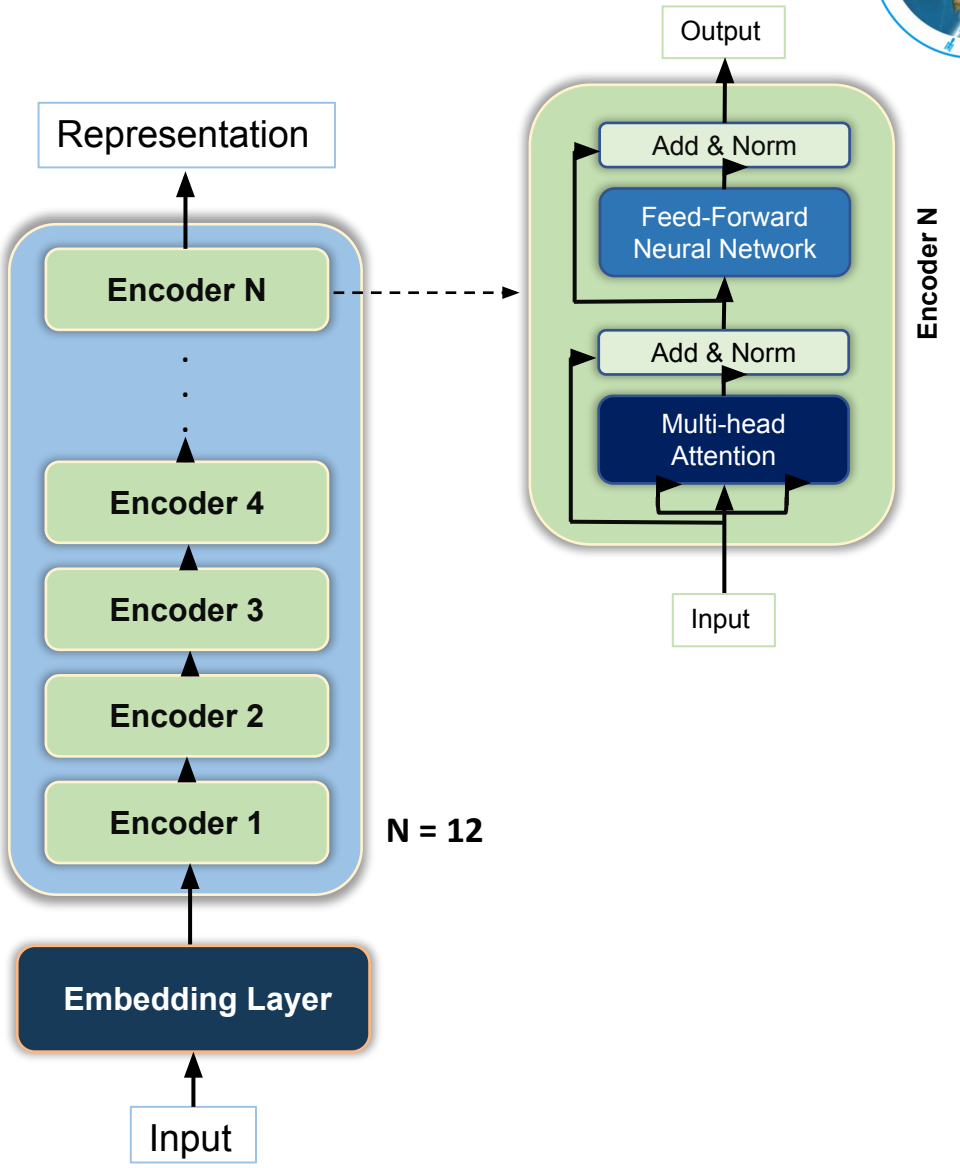
$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)$$



- **Machine Translation**
- **Text Summarization**
- **Fiction Writing**
- **Writing Computer Code and more.**

Overview and Architecture of Google BERT

- BERT Introduction by Google in 2018.
- Revolutionized NLP with bidirectional training.
- Combines Transformer and unsupervised pretraining for context-rich understanding.
- Focuses on contextual understanding with an encoder-only model.
- 12 layers, 768 hidden units, 12 attention heads, and 110M parameters.
- Pioneered contextualized word embeddings, enhancing language model performance.
- Pre-trained on diverse datasets. Effective generalization to a variety of downstream natural language processing tasks.
- Success of BERT laid the groundwork for subsequent language model developments.



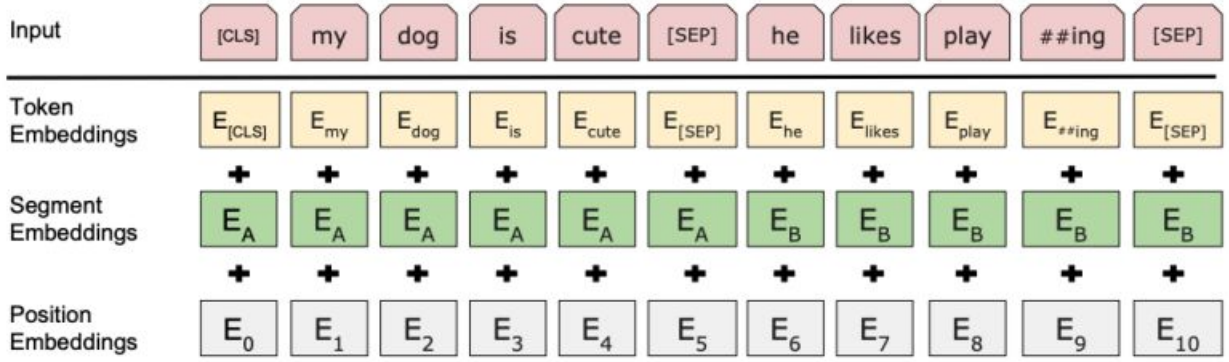
Overview and Architecture of Google BERT Cont..

BERT Pretraining:

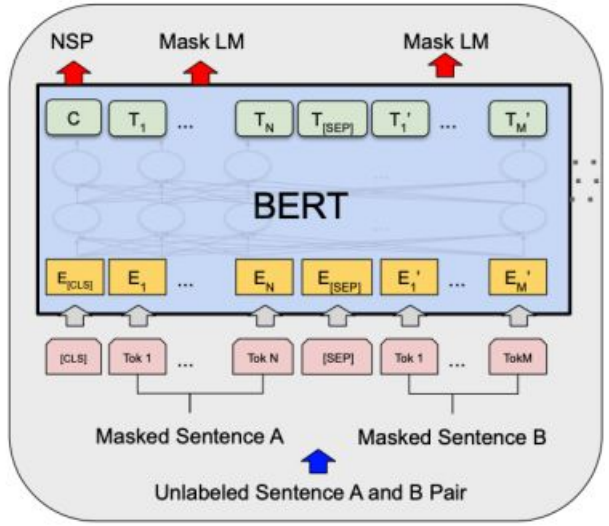
BERT is trained on the BooksCorpus (800M words) and Wikipedia (2,500M words).

BERT is pretrained with a linear combination of two objectives:

Task 1: Masked LM
mask 15% of all WordPiece tokens in each sequence at random
Task 2: Next Sentence Prediction (NSP)
IsNext vs NotNext



BookCorpus and Wikipedia



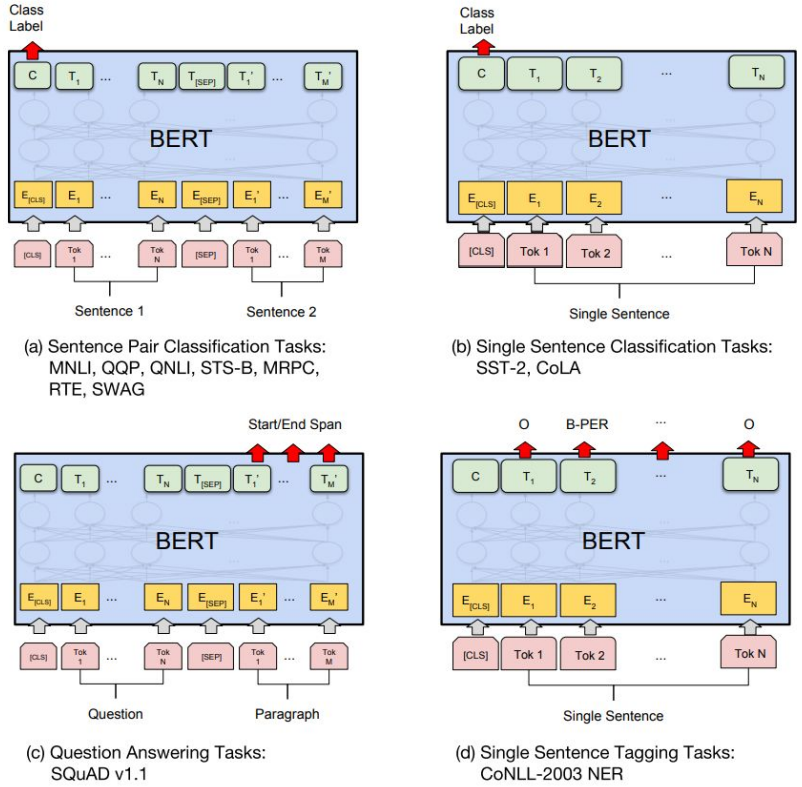
Pre-training

Overview and Architecture of Google BERT Cont..

BERT Finetuning:

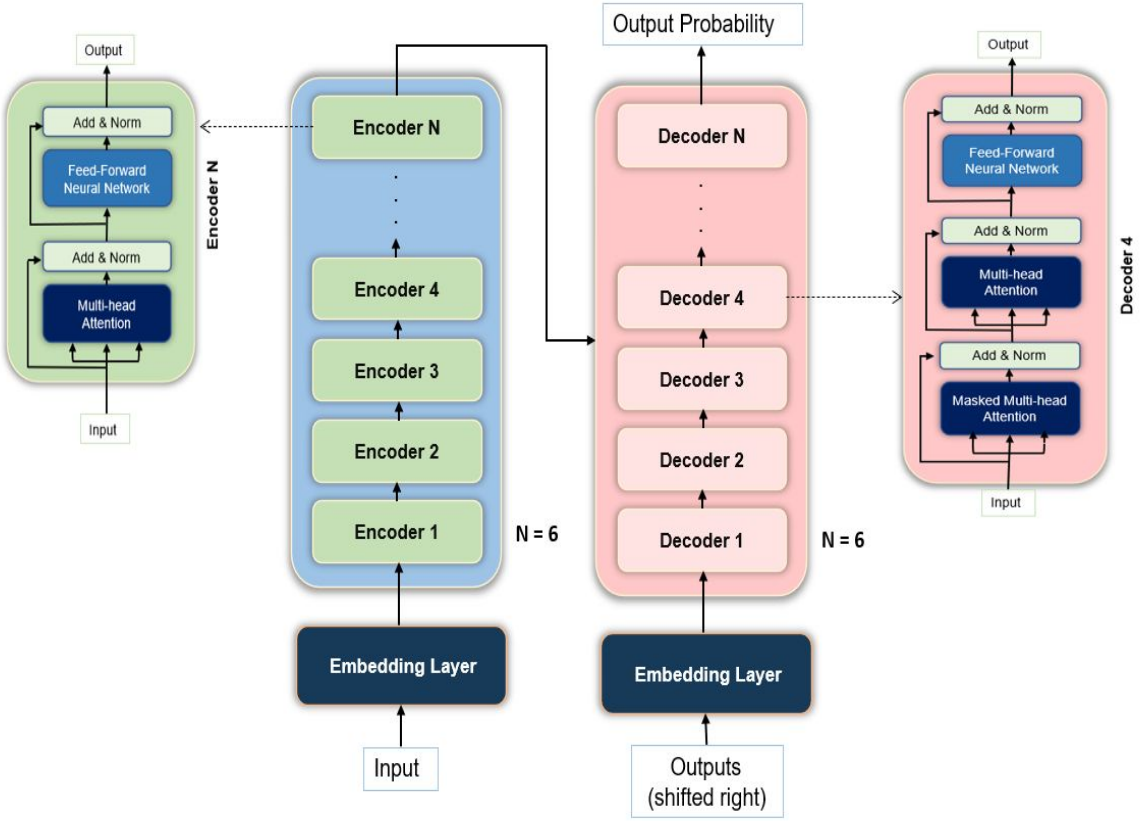
BERT can be fined for various NLP task in supervised manner.

1. Sentence pair classification task.
2. Single sentence classification task.
3. Question Answering task.
4. Single sentence tagging task etc.



Overview and Architecture of Facebook BART

- BART introduced by Facebook AI in 2019.
- It is Encoder-Decoder (seq2seq) Transformer model.
- 12-layer, 768-hidden, 8-heads, 139M parameters.
- Two main parts: a bidirectional encoder (BERT-like) and an autoregressive decoder (GPT-like).
- Pre-trained on diverse data, allowing it to learn rich representations and generalize well to downstream tasks.
- BART is designed to be task-agnostic, making it suitable for a wide range of applications, including summarization, text completion, and more.
- BART has demonstrated state-of-the-art performance in tasks such as text summarization, showcasing its effectiveness in capturing and generating meaningful content.



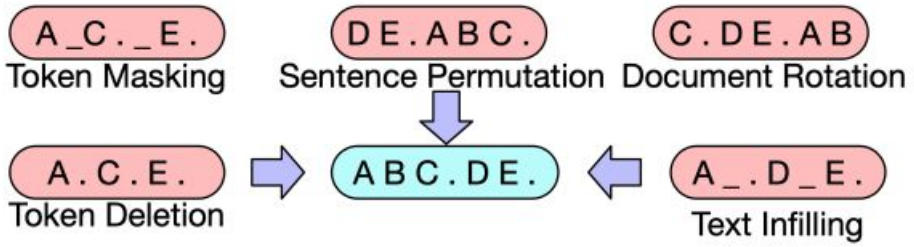
Overview and Architecture of Facebook BART Cont..

BART Pretraining:

- BART is trained by corrupting documents and then developing a model to restore the original text.
- Optimizing a reconstruction loss—the cross-entropy between the decoder's output and the original document.
- **The process is divided into two key components:**

  1. Masked Language Model (MLM) Training
  2. Auto-Regressive Training

BookCorpus and Wikipedia

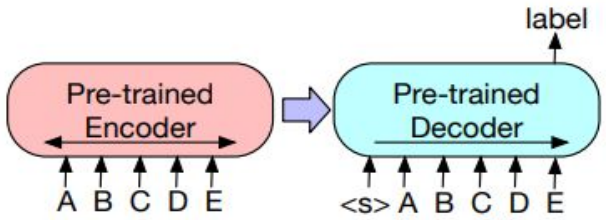


Overview and Architecture of Facebook BART Cont..

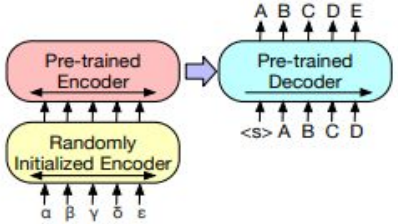
BART Finetuning:

BART can be fined for various NLP task in supervised manner

1. Sequence Classification Tasks.
2. Token Classification Tasks.
3. Sequence Generation Tasks.
4. Machine Translation Tasks.
5. Document Summarization Tasks etc.



Sequence Classification



Machine Translation

Some Open Source Pretrained Models

Model Name	Team	Architecture	Parameter
GPT-2 (Generative Pre-trained Transformer 2)	OpenAI	Transformer Dec.	774M
GPT-3 (Generative Pre-trained Transformer 3)	OpenAI	Transformer Dec.	175B
RoBERTa (Robustly optimized BERT approach)	Facebook AI	Encoder	355M
XLNet (eXtreme Language understanding Network)	Google AI	Encoder	340M
DistilBERT (Distilled Bidirectional Encoder)	Hugging Face	Encoder	66M
T5 (Text-To-Text Transfer Transformer)	Google AI	Transformer	11B
LLaMA2 (Large Language Model Meta AI)	Meta AI	Transformer	70B
LaMDA (Language Model for Dialogue Applications)	Google AI	Transformer Dec.	137B
PaLM (Pathways Language Model)	Google AI	Transformer	540B

Conclusion

Our exploration of transformer architecture in machine translation emphasized its revolutionary impact on efficiency through parallel processing during training and inference. This architectural shift significantly accelerates computational speed. We delved into attention mechanisms, revealing their role in crafting context-based embeddings by focusing on specific elements within input sequences, thereby enhancing contextual understanding.

The journey extended to Transfer Learning in Natural Language Processing (NLP), where we dissected advanced models, uncovering the intricacies of their architectures, pretraining strategies, and fine-tuning approaches. This exploration showcased the transformative influence of transfer learning on various NLP tasks. Additionally, we surveyed open-source models, highlighting the collaborative nature of NLP research.

References

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. OpenAI
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension
- Jay Alammar Blog. Visualizing machine learning one concept at a time. https://jalammar.github.io/
- Jason Brownlee Blog. Machine Learning Mastery. https://machinelearningmastery.com/blog/
- Sudharsan Ravichandiran. Getting Started with Google BERT - Build and train state-of-the-art natural language processing models using BERT

Thank You !

Any Questions ?