Transformers

November 22, 2023

Self-Attention

Multi-head Attention

DO I KNOW MY NEIGHBOURS?

Traditional domain in an windowed (ramp or otherwise) approach

- Frequency count - cooccurence
- Correlation

SVD Domain

- The transformation of the incidence matrix into SVD domain leads to the define the relationship between two words based on the magnitude of the diagonal matrix Σ

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T \tag{1}$$

  The low-rank approximation captures as much information/energy as possible with respect to the word relationships

GloVe Domain

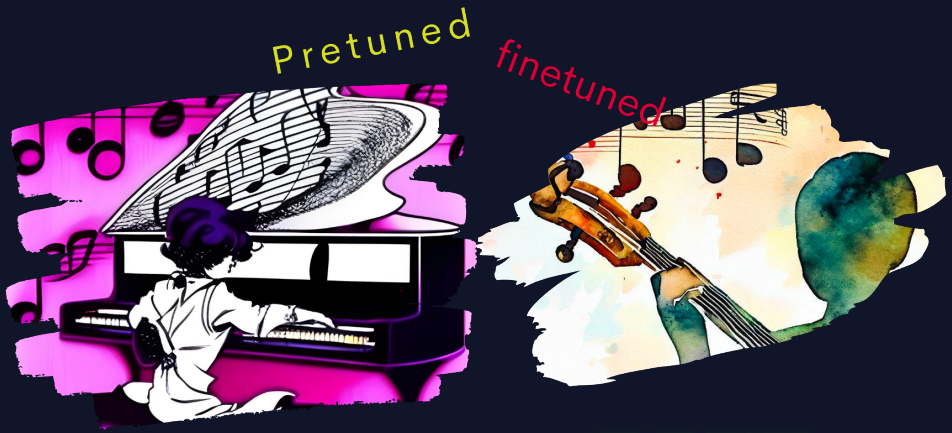
- Ratio of the co-occuring word probabilities are transformed into word embeddings

LANGUAGE MODELS

Useful in

- Auto-generation of sentences
- Auto-completion of sentences
- Creating contextualized word-vectors
- Machine Translation

PRE-TUNED AND FINE-TUNED



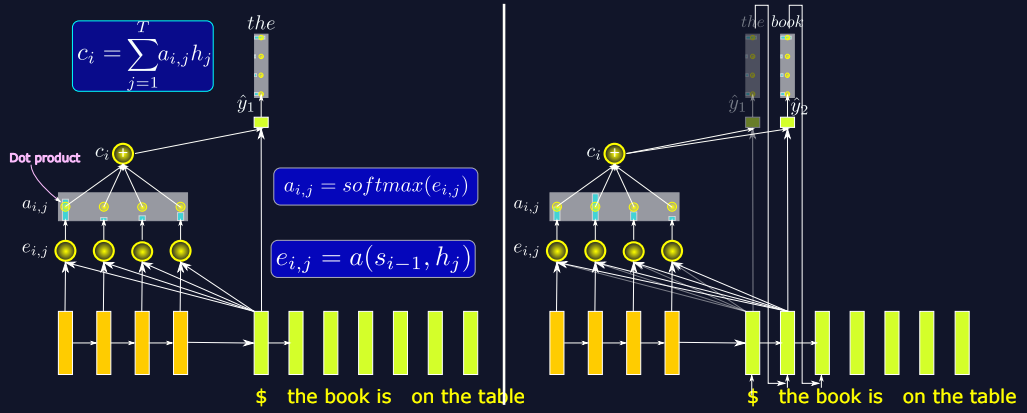
TRANSFER LEARNING

- Models are trained on a specific task - Word-embedding $\rightarrow$ build contextual embedding
- Use the learned weights for another task - attention in machine learning - transfer the hidden layer (contextual information) and use it for another task - mostly done using bidirectional LSTM - Embeddings from Language Model (ELMO)
  - Train using a sentence - forward and reverse Sequence
  - Concatenate hidden layers of (1)
  - Depending on the task on hand, multiply each (2) vector by a scale factor
  - Sum (2) to get the contextualized embedding

  $\rightarrow \mathrm{ELMO}_k = \gamma \sum_{j=0}^{L} \mathrm{SoftMax}_j h_{j,k}$

  where $h_{j,k} = \mathrm{BiLSTM}(w_{1:T}, k)$, k is the index of the word and $1 \leq j \leq L$

ATTENTION



$e_{i,j}$−attention score

$a_{i,j}$−attention distribution

c_i−attention output

DID I KNOW MORE ABOUT MYSELF AND MY NEIGHBORS?

- How much should I be influenced by my neighbors?
- How do I capture information about my neighbors?
- Ideally a model creates a word vector that has enough information about its neighbors irrespective of how well it is connected with them statistically

ANY CHANGE REQUIRED?

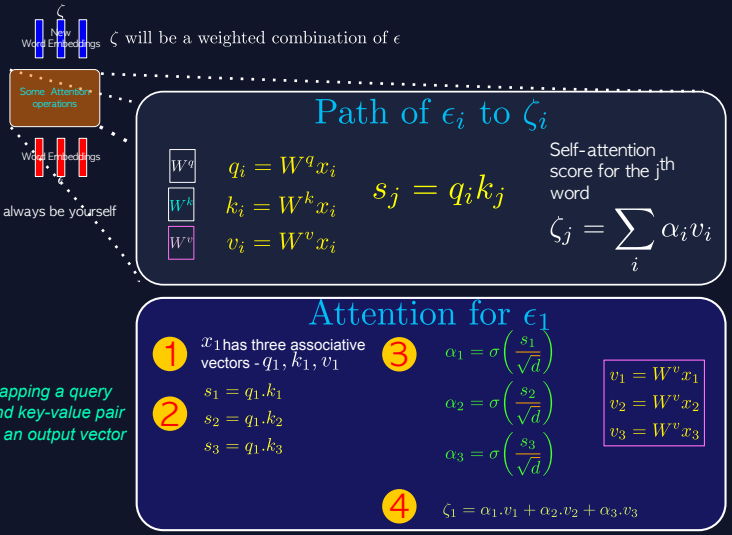
- Is it possible to encode the time-series differently?
- How do I reduce the complexity of recurrence?
- Is it possible to look at each embedding differently?

IDEAL ENCODING OF EMBEDDING

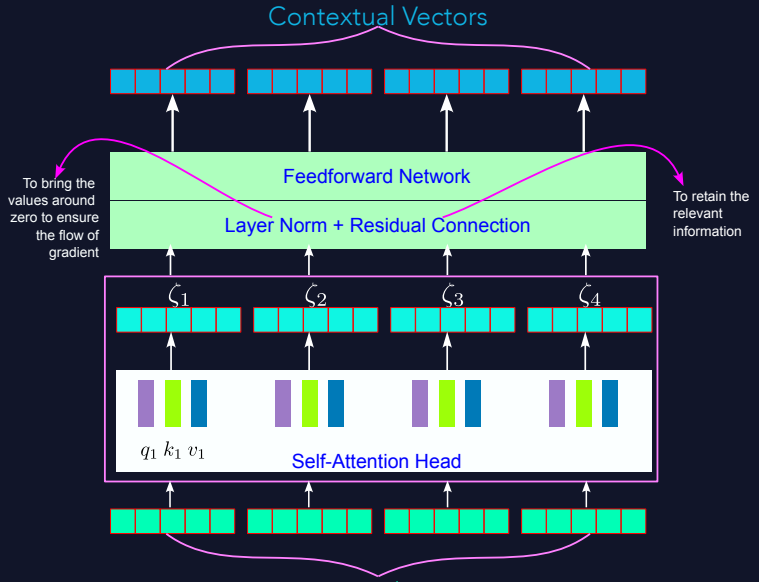
ζ will be the weighted combinations of ϵ

Given a set of vector values, and a vector query, attention is a technique to compute a weighted sum of the values, dependent on the query

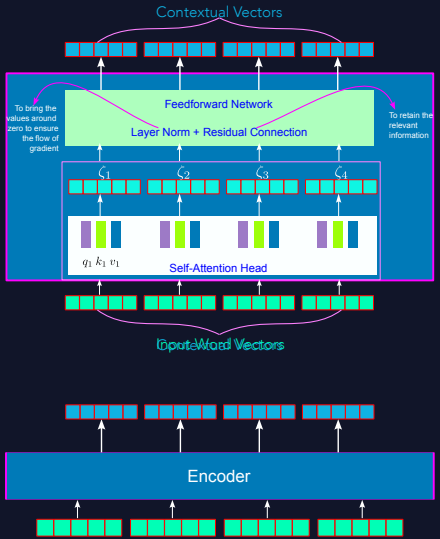
SELF-ATTENTION LAYER

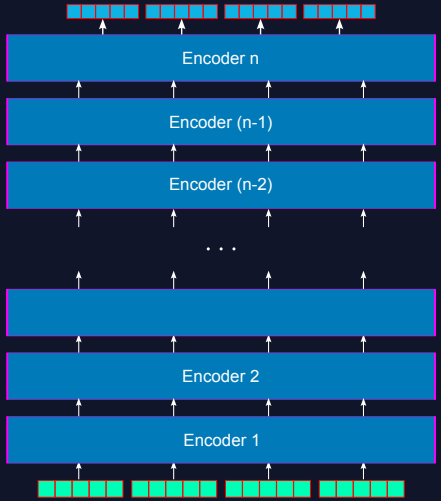


SINGLE TRANSFORMER LAYER



TRANSFORMER ENCODER





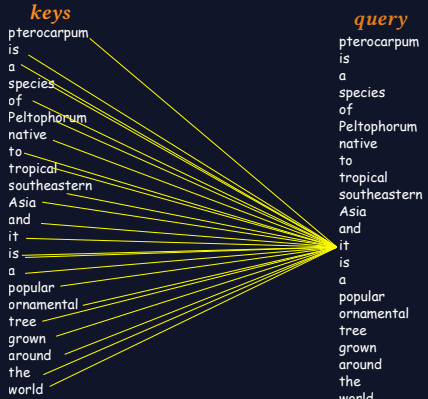
SELF-ATTENTION

Peltophorum pterocarpum is a species of **Peltophorum**, native to tropical southeastern Asia and **it** is a popular ornamental tree grown around the world

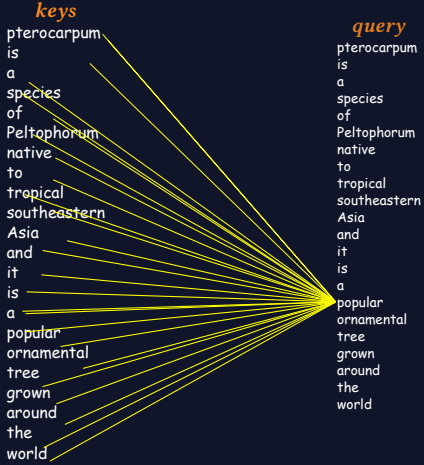
What does it refer to?

Self-attention allows us to associate *it* with Peltophorum pterocarpum

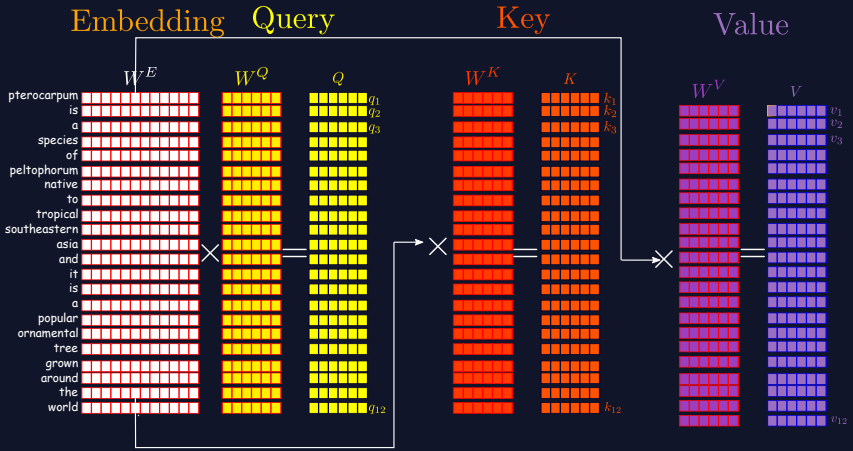
SELF-ATTENTION - WORD LEVEL



Self-attention allows each word to align itself to other words using their positions and looks for clues for a better contextul encoding

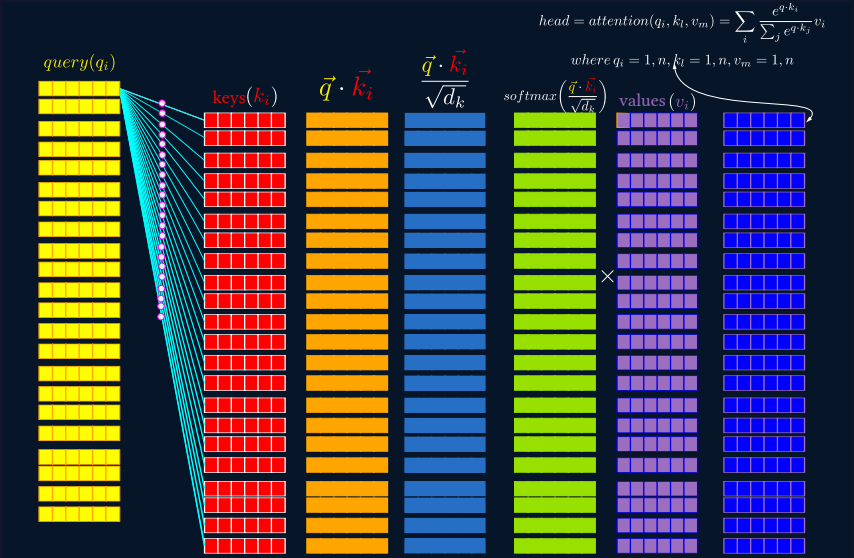


CREATING QUERY, KEY AND VALUES



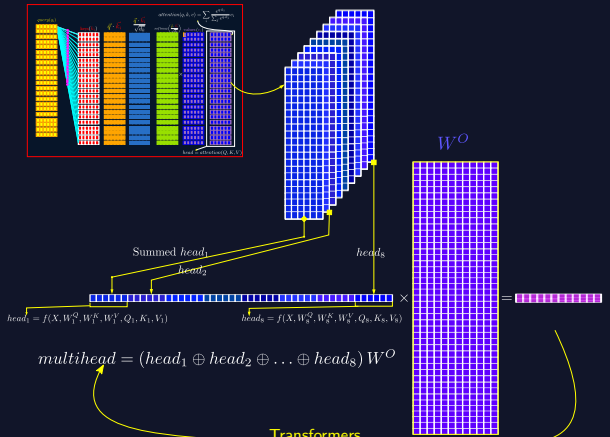
W^Q, W^K and W^V are learned during the training process

COMPUTING ATTENTION(Q,K,V)

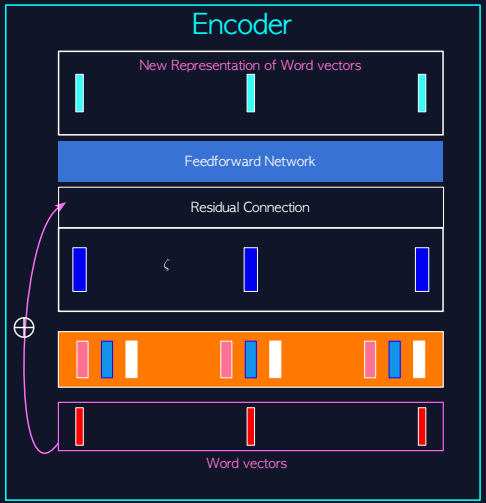


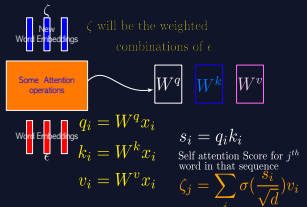
MULTI-HEAD ATTENTION

The summed attention score, $attention(q, k, v) = \sum_i \frac{e^{q \cdot k_i}}{\sum_j e^{q \cdot k_j}} v_i$, has information about all the word, but may have more information about words that have similarity scores higher than others in that context.

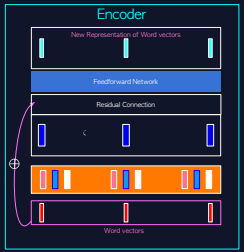


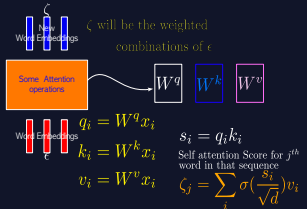
SELF-ATTENTION

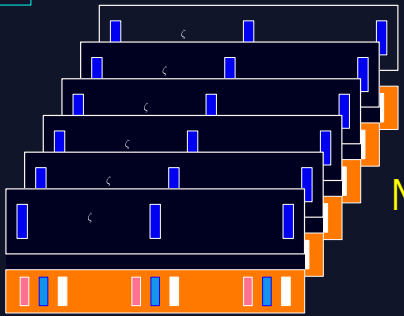




MULTIHEAD ATTENTION







POSITIONAL ENCODING

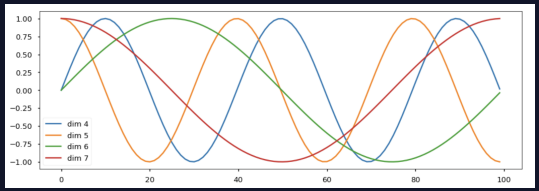
- RNN models encode the time signal in a sequential manner
- Positional encoding is important for contextual learning
- Transformers a separate positional vector is added to the embedding
- Positional encoding should be unique, not just a scalar
- positional values should be bounded
- Positional values between any two positions should be consistent
- Sentence length should not be a constraint
- Deterministic

POSITIONAL ENCODING

$$\vec{pe}_t^{\,i} = \begin{cases} \sin(\omega_k \cdot t), & \text{if } i = 2k \\ \cos(\omega_k \cdot t), & \text{if } i = 2k+1 \end{cases}$$

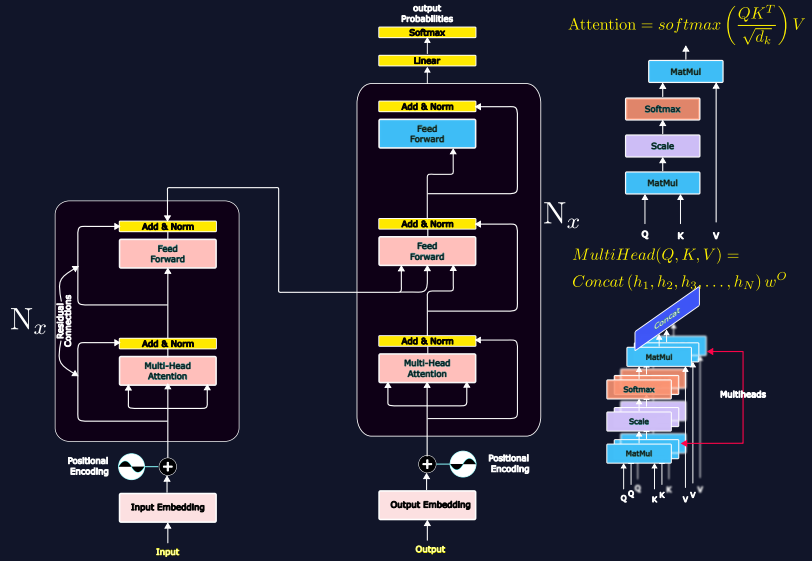
where $\omega_k = \dfrac{1}{10000^{(i/d)}}$.

This positional vector will be a pairs of sin and cos values and i is the index of the elements in the vector



1. Good reference for Transformer - The Illustrated Transformer, Jay Alammar
2. The Annotated Transformer - Austin Huang et al

TRANSFORMER ARCHITECTURE



NUMBER OF PARAMETERS IN A TRANSFORMER

Embedding Layer - $N_{emb} = V \times D$
where N_{emb} is the number of parameters in the embedding layer, V is the vocabulary size and D is the embedding dimension

Multi-Head Attention -
$N_{attn} = 3 \times D \times d_k \times h + h \times d_v \times d_k + 2 \times h \times d_k (biasses)$
where N_{attn} is the number of parameters in multi-head attention, D is the embedding dimension, d_k is the dimension of the key vector, h is the number of heads and d_v is the dimension of the value vector

Feed-Forward Network -
$N_{ffn} = 2 \times D \times H + H + H \times D$
where N_{ffn} is the number of parameters in the feed-forward network, D is the embedding dimension and H is the hidden dimension

Layer Norm - $N_{ln} = 2 \times D$
where N_{ln} is the number of parameters in the layer normalization, D is the embedding dimension

Total Number of Parameters per Encoder/Decoder Layer -
$N_{layer} = N_{attn} + N_{ffn} + N_{ln}$

Total Number of Parameters in the Transformer - $N_{total} = N_{emb} + N_{layer} \times L$
where L is the number of layers

WORD EMBEDDING

1. Word embedding is fundamental to Deep Learning of NLP
2. Contextual word embedding is now used in most of the downstream applications

CLOZE TEST

According to a report in yesterday's newspaper (1)——— police dog was taken to Raj Bhavan (2)——— Monday. This was to trace the (3)——— of the "very important horse" which (4)——— reported missing on Sunday. The dog picked (5)——— the scent on some traces of (6)——— and ran a few yards before losing the (7)———. The police have launched a vigorous (8)——— into the whole affair. They have (9)——— the services of a forensic expert, (10)——— fingerprint expert and a photographer. (11)——— are now fourteen horses at Raj Bhavan (12)——— are kept in a large shed near the gate.

1	once	a	new
2	at	next	on
3	police	killers	dogs
...	...	...	...
11	There	We	So
12	who	were	which

The purpose of the Cloze test is to measure the reading comprehension of a student with respect to grammar, usage, vocabulary, and contextual understanding.

MASKED LM

Objectives

- Predict the masked word using the context
- Combine left and right contexts to predict the masked word
- Use sentence pairs to predict the next sentence
- Use the trained model for token-level and sentence-level tasks
- Mask some of the tokens from the input
- Predict the original token using its context
- Use Bidirectional deep transformer model fuse the left and right context and develop a contextual representation

HOW DO WE MASK?

- 15% of the words are randomly masked
- Perform the following operation on the the i^{th} token
  1. Replace with a $<\mathcal{MASK}>$ 80% of the time
  2. Replace with a random token 10% of the time
  3. Retain the original 10% of the time

BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) - ARCHITECTURE

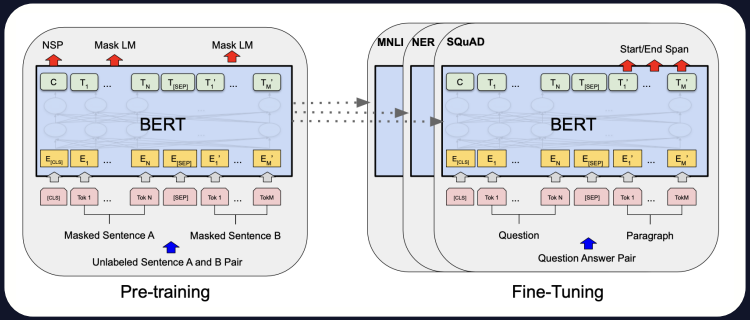


Figure: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architec- tures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

GENERAL PURPOSE TRANSFORMER - GPT

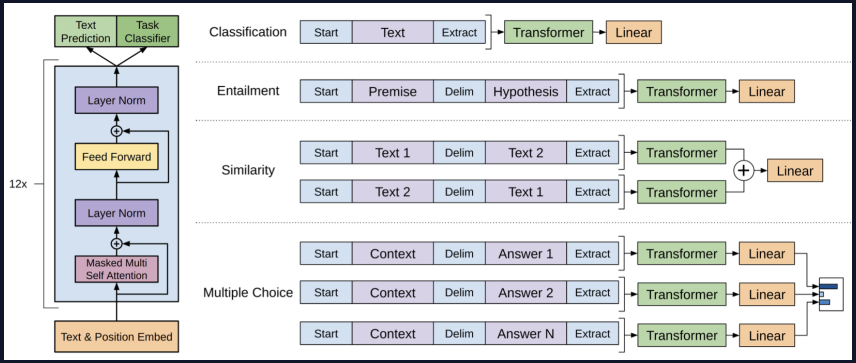
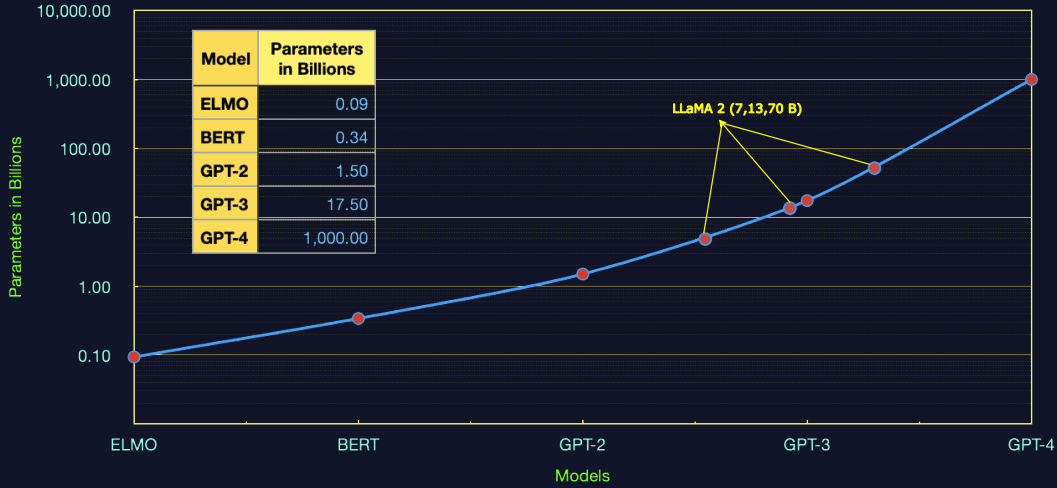


Figure: (left) Transformer architecture and training objectives used in this work. (right) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer[1]

[1] GPT

PARAMETERS OF LARGE LANGUAGE MODEL



Model	Parameters in Billions
ELMO	0.09
BERT	0.34
GPT-2	1.50
GPT-3	17.50
GPT-4	1,000.00

SAMPLE TEXT GENERATION - MINIGPT

```
const LEARNING_RATE: f64 = 0.0003;
const BLOCK_SIZE: i64 = 128;
const BATCH_SIZE: i64 = 64;
const EPOCHS: i64 = 100;
const SAMPLING_LEN: i64 = 4096;
```

```
Config {
vocab_size: 130,
n_embd: 128,
n_head: 8,
n_layer: 8,
block_size: 128,
attn_pdrop: 0.1,
resid_pdrop: 0.1,
embd_pdrop: 0.1,
};
```

for the gbr genome determines a comprehensive comparison from existing quality was commonly important to a lung w fusion if they can be a compared x-rayed, with the blood scenario of the lesion of t-cell support probability.in this later, the next for the first room of responses to materials and infection has classified a retrospective response of younger genes [138] .