Hello everyone, this video is dedicated to the implementation aspect of the Transformer using the Python language and the PyTorch framework.

In the last video, we have discussed all building blocks of transformer, the model proposed in 2017 by Vaswani and his colleagues in the paper “Attention is all you need”. This model has revolutionized not only the field of Natural Language Processing but also the whole world of artificial intelligence.

In this video we will talk about the implementation part of this model. So for preparing a good boarding for this long video, if it is not the case, I recommend you to check my previous video about transformer to have a good understanding of all building blocks of the transformer.

**What we will obtain from this video includes:**

* Building a Transformer package that contains the source code to build Transformer blocks, utility functions to process the raw data, etc.
* This video will also demonstrate how to write the training and inference scripts in just a few lines of code, which call the functions and classes from the Transformer package.
* Moreover, along with the code development, I always provide the notebook files to experiment with the code and to verify that it works.

Well, if you are ready, let’s jump in.

---

Well, we will look at our plan for transformer implementation.

There are two main parts.

Firstly, we will be implementing the training process of the transformer model. In the case of transformer, this process takes source sentences in one language and the target sentences in another language as inputs, then it processes the training to obtain the trained model that can create a target sentence, given a source sentence.

Secondly, when the trained model is obtained, we will implement and test the inference part. At this step, given we have a sentence in one language like English, add it as input of the trained model then we obtain at the output of the trained model the translated sentence in another language like French.

---

What are the implementation objectives.

First, the implementation should help you to see the whole picture.

Then it walk you through the detail map to go from A to Z in the implementation of transformer?

Another important aspect is clean code? Which involves: well-structured source code, modular, well-documented and testable.

Obtaining all of the objectives makes the implementation easy to understand, easy to maintain for debugging and easy to scale/modify/customize the source code when needed.

---

What is my approach to obtain all the objectives and give a good understanding of transformer implementation?

My formular is that.

First, I will show you the source code structure in the Python package template, which file contains which function, which function calls which class, etc.

After that, I will walk you through coding all Python functions and class of each Transformer’s block step by step.

Then I will present the notebooks playground, presenting a concrete usage example of that function or class,

And finally showing you the expected outputs to validate that the function works properly.

Combining the four elements allows us to obtain a good understanding of transformer implementation.

At this point, we are ready to jump into the detail part.

---

Now I am going to present the training part, which is the process allows the model to learn features from data, for example the case of machine translation task, it learns feature from source text and target text, then it adjusts its weights values so that the model can generate the text as same as the target text, given a source text.

--

Now I am going to talk about the two main steps in training processing:

First, we load the configuration. It is a Json file containing a Python Dictionary, where the keys are name of different parameters serving different steps in training and inference, for example the model building or data loading, data split etc. The values can be string or numerical values.

Then, we go into the training engine, which includes 3 main step: pre-processing data, create and initiate transformer model, and training loop. We will dive deeper into each of the sub-step.

---

For preprocessing data, we will prepare the functions like get\_dataset, get\_tokenizer, create encoder mask, padding mask, causal mask, decoder mask and the class like DataPreprocessor. We also use the class DataLoader from PyTorch framework.

---

For building a transformer model, we need to build the classes: InputEmbedding, PositionalEncoding, Encoder, Decoder, and Projection. For Encoder and Decoder, we need to build EncoderLayer and DecoderLayer respectively. The latter is derived from the atomic classes: MultiHeadAttention, FeedForward, ResidualConnection, LayerNormalization.

---

For training loop, we need to prepare loss function, optimizer, we also need to prepare evaluation\_step and model\_inference function for validation stage during training.

---

Now put it all together to see the whole picture. It seems quite complicated to successfully run the training process of transformer but break the complex task into small modular functions and classes, we can easily move on and turn the complex thing into multiple simple sub-component. Well now you have high level and structured overview of what will be implemented, let’s jump into the exciting parts.

---

Now we look at source code tree structure. I will build transformer source code as a Python package and leverage PyTorch framework to build Transformer model.

For this, we create a folder: transformer. Inside this package, there are 3 main groups:

* \_\_init\_\_.py file, this is a default file to let Python know that it’s parent folder is a Python package
* Then, Python class
* And the third group is Python functions

---

For Transformer class, we will create the following python files: data, layer, encoder, decoder, and model.

In which,

* data file contains DataPreprocessor class
* layer file contains all classes to create building blocks of transformer: Input Embedding, PositionalEncoding, MultiHeadAttention, FeedForward, LayerNorm, ResidualConnection, Projection
* For encoder file, we have the classes: Encoder Layer and Encoder
* For decoder file, we have the classes: Decoder Layer and Decoder
* For model file, we have Transformer class

---

For Transformer functions, we will create the following Python files: mask, utils and engine.

In which,

* Mask file contains functions to create encoder mask, padding mask, causal mask and decoder mask
* Utils file contains functions to load configuration, get dataset, get tokenizer, prpeprocessing data and create transformer model
* Finally, engine files contains functions to call the training, evaluation during training and inference after training

---

Now we will see how to implement the class and functions for the data preparation steps.

We will firstly talk about the function to load the configuration, to get the dataset, containing pairs of English and French sentence, the function to get the source and target tokenizer.

After that the functions to create the encoder, padding, causal and decoder masks,

Then the Data PreProcessor class.

And finally the function preprocessing data.

Let’s jump into the excited part.

---

**Get Config**

This Python function load\_config is used to load a configuration file in JSON format.

Here's a step-by-step explanation:

1. The function takes one argument config\_file\_path, which is the path to the configuration file.
2. It uses Python's built-in open function to open the file in read mode ("r").
3. The with statement is used here which allows for clean resource management, automatically closing the file after the nested block of code is executed.
4. The json.load function is used to parse the JSON file and convert it into a Python dictionary. This dictionary is stored in the config variable.
5. The function then returns this config dictionary which contains the data from the JSON configuration file. This can be used in the rest of the program to access configuration values.

---

**Get Config - Notebook**

Now, in the jupyter notebook file, we can call the load\_config function, add the input parameter as the path to the config.json file, then the output is assigned to the variable config as a Python dictionary type.

Afterwards, we can use access to the key and value in the configuration data.

For example, we can print some keys and values like: source language, target language, batch size, number of epochs like this.

Finally, we can see what results we can obtain at the end.

Great! We can now load and access to the configuration data, to use it for the whole program later.

---

**Get Dataset**

This function get\_dataset is used to load a dataset for a machine translation task. Here's a step-by-step explanation:

1. Dataset name and language pair: The function first retrieves the name of the dataset and the source and target languages from the configuration. It then constructs the language pair string by concatenating the source and target languages with a hyphen.
2. Split: It retrieves the dataset split (such as "train", "validation", or "test") from the configuration.
3. Load dataset: It uses the load\_dataset function from the datasets library of Hugging Face to load the specified split of the dataset on the Hugging Face Hub The load\_dataset function takes the name of the dataset and the language pair as arguments. This function returns a DatasetDict object, which is a dictionary-like object that maps the split names to the corresponding Dataset objects.
4. Return dataset: Finally, the function returns the loaded dataset. This dataset can be used to train and evaluate a machine translation model.

**Get Dataset - Notebook**

This script is used to load a dataset for a machine translation task using a Transformer model. Here's a step-by-step explanation:

1. Import necessary modules: The script starts by importing necessary modules and functions. It also appends the parent directory to the system path to access modules in different directories.
2. Load configuration: The load\_config function is used to load the configuration settings from a JSON file. The path to the configuration file is specified as ../config.json.
3. Load dataset: The get\_dataset function is used to load the dataset based on the configuration settings. This function uses the load\_dataset function from the datasets library to load the specified dataset.
4. Print first two translations: The script then prints the first two items from the 'translation' field of the dataset. This field typically contains the source and target sentences for the machine translation task.

In summary, this script is a useful tool for loading and inspecting the dataset used in a machine translation task.

**Get Tokenizer**

This function get\_tokenizer is used to load a tokenizer for a specific language. If the tokenizer already exists (i.e., it has been saved to a file), it loads the tokenizer from the file. If the tokenizer does not exist, it creates a new one, trains it on the dataset, and then saves it to a file. Here's a step-by-step explanation:

1. **Tokenizer name and path**: The function first retrieves the name of the tokenizer from the configuration and constructs the path to the tokenizer file by appending the language to the tokenizer name.
2. **Load tokenizer**: If the tokenizer file exists, it loads the tokenizer from the file using the Tokenizer.from\_file method.
3. **Create and train tokenizer**: If the tokenizer file does not exist, it creates a new tokenizer with a word-level model and an unknown token ([UNK]). It also sets the pre-tokenizer to split the text into words based on whitespace.

It then creates a trainer with a minimum frequency of 2 and special tokens for padding ([PAD]), unknown words ([UNK]), start of sentence ([SOS]), and end of sentence ([EOS]). The tokenizer is then trained on the dataset using the train\_from\_iterator method. The dataset is assumed to be a dictionary with a key "translation" that contains a list of dictionaries, each with a key for each language and the corresponding text as the value.

1. **Save tokenizer**: After training the tokenizer, it saves the tokenizer to a file so that it can be loaded quickly in the future.
2. **Return tokenizer**: Finally, the function returns the tokenizer. This tokenizer can be used to convert text to tokens and vice versa.

**Tokenizer Notebook**

This script is used to load and explore the properties of source and target tokenizers in a machine translation task using a Transformer model. Here's a step-by-step explanation:

1. Import necessary modules: The script starts by importing necessary modules and functions. It also appends the parent directory to the system path to access modules in different directories.
2. Load configuration: The load\_config function is used to load the configuration settings from a JSON file.
3. Load dataset: The get\_dataset function is used to load the dataset based on the configuration settings.
4. Load tokenizers: The get\_tokenizer function is used to load the source and target tokenizers. These tokenizers are used to convert text to tokens and vice versa.
5. Explore tokenizer properties: The script then prints out various properties of the tokenizers, such as the vocabulary size, how they encode text and decode token,

In summary, this script is a useful tool for exploring the properties of the tokenizers used in a machine translation task.

**Create masks**

This code snippet contains four functions that are used to create masks for a transformer model in PyTorch. These masks are used to prevent the model from looking at certain parts of the input data.

1. create\_encoder\_mask: This function creates a mask for the encoder inputs. It takes the encoder input IDs and the ID of the padding token as arguments. The mask has the same shape as the encoder input IDs, and contains a 1 at each position where the corresponding input ID is not a padding token, and a 0 where it is.
2. create\_padding\_mask: This function creates a padding mask for the decoder inputs. It works similarly to create\_encoder\_mask, but does not add extra dimensions to the mask.
3. create\_causal\_mask: This function creates a causal mask, which is used to ensure that each position in the decoder can only attend to earlier positions in the sequence. It takes the sequence length as an argument and creates a square mask of that size, with 1s below and on the diagonal and 0s above the diagonal.
4. create\_decoder\_mask: This function creates a mask for the decoder inputs by combining a padding mask and a causal mask. It takes the decoder input IDs, the ID of the padding token, and the sequence length as arguments. The resulting mask contains a 1 at each position where the corresponding input ID is not a padding token and the position is not ahead in the sequence, and a 0 otherwise.

**Playground – Create encoder mask**

This Python script is used to create and print an encoder mask for a given input sequence using a transformer model.

Here's a step-by-step explanation:

1. The script first modifies the system path to include the parent directory (../). This is done so that Python can import modules from files in that directory.
2. It then imports the PyTorch library and all functions from the transformer.mask module.
3. The script creates a tensor en\_input\_ids representing the input sequence to the encoder. Each number in the tensor corresponds to the ID of a token in the sequence. The pad\_id tensor is also created, representing the ID of the padding token.
4. The create\_encoder\_mask function from the transformer.mask module is called with en\_input\_ids and pad\_id as arguments. This function returns a mask that indicates where the input sequence has tokens (value 1) and where it has padding (value 0). The mask has the same length as the input sequence and is reshaped to have two extra dimensions at the start.
5. The script then prints the input IDs, their shape, the encoder mask, and its shape. The shape of a tensor is a tuple indicating the size of each dimension. For the input IDs, the shape should be (seq\_len,), where seq\_len is the length of the sequence. For the encoder mask, the shape should be (1, 1, seq\_len).

**Playground – Create padding mask**

This Python script is used to create and print a padding mask for a given input sequence using a transformer model.

Here's a step-by-step explanation:

1. The script first modifies the system path to include the parent directory (../). This is done so that Python can import modules from files in that directory.
2. It then imports the PyTorch library and all functions from the transformer.mask module.
3. The script creates a tensor de\_input\_ids representing the input sequence to the decoder. Each number in the tensor corresponds to the ID of a token in the sequence. The pad\_id tensor is also created, representing the ID of the padding token. The seq\_len variable is set to the length of the sequence.
4. The create\_padding\_mask function from the transformer.mask module is called with de\_input\_ids and pad\_id as arguments. This function returns a mask that indicates where the input sequence has tokens (value 1) and where it has padding (value 0). The mask has the same length as the input sequence.
5. The script then prints the input IDs, their shape, the padding mask, and its shape. The shape of a tensor is a tuple indicating the size of each dimension. For the input IDs, the shape should be (seq\_len,), where seq\_len is the length of the sequence. For the padding mask, the shape should be (1, seq\_len).

**Playground – Create causal mask**

This Python script is used to create and print a causal mask for a given input sequence using a transformer model.

Here's a step-by-step explanation:

1. The script first modifies the system path to include the parent directory (../). This is done so that Python can import modules from files in that directory.
2. It then imports the PyTorch library and all functions from the transformer.mask module.
3. The script creates a tensor de\_input\_ids representing the input sequence to the decoder. Each number in the tensor corresponds to the ID of a token in the sequence. The pad\_id tensor is also created, representing the ID of the padding token. The seq\_len variable is set to the length of the sequence.
4. The create\_causal\_mask function from the transformer.mask module is called with seq\_len as an argument. This function returns a mask that ensures that each position in the decoder can only attend to earlier positions in the sequence. It creates a square mask of size seq\_len with 1s below and on the diagonal and 0s above the diagonal.
5. The script then prints the input IDs, their shape, the causal mask, and its shape. The shape of a tensor is a tuple indicating the size of each dimension. For the input IDs, the shape should be (seq\_len,), where seq\_len is the length of the sequence. For the causal mask, the shape should be (1, seq\_len, seq\_len).

**Playground – Create decoder mask**

This Python script is used to create and print a decoder mask for a given input sequence using a transformer model.

Here's a step-by-step explanation:

1. The script first modifies the system path to include the parent directory (../). This is done so that Python can import modules from files in that directory.
2. It then imports the PyTorch library and all functions from the transformer.mask module.
3. The script creates a tensor de\_input\_ids representing the input sequence to the decoder. Each number in the tensor corresponds to the ID of a token in the sequence. The pad\_id tensor is also created, representing the ID of the padding token. The seq\_len variable is set to the length of the sequence.
4. The create\_decoder\_mask function from the transformer.mask module is called with de\_input\_ids, pad\_id, and seq\_len as arguments. This function returns a mask that combines a padding mask and a causal mask. The padding mask indicates where the input sequence has tokens (value 1) and where it has padding (value 0). The causal mask ensures that each position in the decoder can only attend to earlier positions in the sequence.
5. The script then prints the input IDs, their shape, the decoder mask, and its shape. The shape of a tensor is a tuple indicating the size of each dimension. For the input IDs, the shape should be (seq\_len,), where seq\_len is the length of the sequence. For the decoder mask, the shape should be (1, seq\_len, seq\_len).

**Data Preprocessor Class**

This Python script defines a class DataPreprocessor that inherits from PyTorch's Dataset class. The purpose of this class is to preprocess data for a transformer model.

Here's a step-by-step explanation:

1. The \_\_init\_\_ method initializes the class with the dataset, source and target tokenizers, source and target languages, and the maximum sequence length. It also retrieves the token IDs for the special tokens [SOS], [EOS], and [PAD].
2. The \_\_len\_\_ method returns the number of items in the dataset.
3. The \_\_getitem\_\_ method preprocesses the data for the item at the given index. It performs the following steps:
   * Retrieves the source and target texts from the dataset.
   * Tokenizes the source and target texts.
   * Calculates the number of padding tokens needed for the encoder and decoder inputs.
   * Creates the encoder and decoder inputs and the decoder target. The inputs start with the [SOS] token, followed by the tokenized text, the [EOS] token (for the encoder input), and the necessary number of [PAD] tokens. The decoder target is the tokenized text followed by the [EOS] token and the necessary number of [PAD] tokens.
   * Creates the masks for the encoder and decoder inputs using the create\_encoder\_mask and create\_decoder\_mask functions. These masks are used in the transformer model to ignore the padding tokens during the attention calculations.
   * Returns a dictionary containing the preprocessed data.

This class is used to create a PyTorch DataLoader that provides batches of preprocessed data to the transformer model during training.

**Data Preprocessing**

The preprocessing\_data function is used to prepare the data for training, validation, and testing of a Transformer model in a machine translation task. Here's a step-by-step explanation:

1. **Get Dataset**: It calls the get\_dataset function to retrieve the raw dataset based on the configuration.
2. **Tokenization**: It calls the get\_tokenizer function to get the source and target tokenizers. These tokenizers are used to convert the source and target texts into sequences of token IDs.
3. **Split Dataset**: It splits the raw dataset into training, validation, and testing datasets. The split is 70% for training, 20% for validation, and 10% for testing.
4. **Data Preprocessing**: It uses the DataPreprocessor class (explained in the previous question) to convert the raw datasets into a format that can be directly used by the Transformer model. This includes tokenization, addition of special tokens, and creation of masks.
5. **Data Loading**: It creates PyTorch DataLoader objects for the training, validation, and testing datasets. These DataLoader objects are used to iterate over the datasets in batches. The batch size for the training dataset is specified in the configuration, while the batch size for the validation and testing datasets is set to 1. The data is shuffled to ensure that the model is not biased by the order of the examples.
6. **Return**: It returns the DataLoader objects for the training, validation, and testing datasets, as well as the source and target tokenizers.

This function is typically called at the beginning of the training script to prepare the data for the training loop.

---

Now we are going to implement building block of transformer under the form of Python class, in which we will talk about Input Embedding, Positional Encoding, Multi-Head-Attention, Feed-Forward, Layer Normalization, Residual Connection, and Projection layer

---

**Input Embedding**

This InputEmbedding class is a part of the Transformer model architecture, specifically used for embedding the input tokens.

Here's a breakdown of the class:

It inherits from nn.Module, which is the base class for all neural network modules in PyTorch.

The \_\_init\_\_ method initializes the class. It takes two parameters: vocab\_size and d\_model. vocab\_size is the size of the vocabulary, and d\_model is the dimension of the embeddings. Inside this method, an embedding layer is created using nn.Embedding with vocab\_size and d\_model as parameters. The d\_model is also stored as an instance variable.

The forward method defines the forward pass of the module. It takes an input x which is expected to be a 2D tensor with shape [batch\_size, seq\_len], where batch\_size is the number of sequences in a batch and seq\_len is the length of each sequence. The method returns the embedded input, scaled by the square root of d\_model. The scaling is a part of the original Transformer model, which helps in stabilizing the gradients.

**Input Embedding – Notebook**

This use case demonstrates how to use the InputEmbedding class to convert sequences of word indices into continuous vector representations (embeddings).

Here's a step-by-step explanation:

1. Import the necessary modules: PyTorch and the InputEmbedding class.
2. Define the vocabulary size (vocab\_size) and the embedding dimension (d\_model). In this case, the vocabulary size is 5000 (i.e., there are 5000 unique words in the vocabulary), and the embedding dimension is 300 (i.e., each word will be represented as a 300-dimensional vector).
3. Create an instance of the InputEmbedding class, passing the vocabulary size and the embedding dimension as arguments.
4. Define a batch of 2 sequences, each sequence containing 4 words. The words are represented by their indices in the vocabulary. This is done using a PyTorch tensor.
5. Pass the sequences through the InputEmbedding instance. This converts the word indices into their corresponding embeddings and scales them by the square root of d\_model.
6. Print the shape of the resulting tensor. The output shape is [2, 4, 300], which corresponds to [batch\_size, seq\_len, d\_model]. This means there are 2 sequences in the batch, each sequence has 4 words, and each word is represented as a 300-dimensional vector.

**Positional Encoding**

The PositionalEncoding class is a part of the Transformer model architecture, specifically used for adding positional information to the input embeddings. This is necessary because the Transformer does not have any inherent notion of the position of words in a sequence.

Here's a breakdown of the class:

It inherits from nn.Module, which is the base class for all neural network modules in PyTorch.

The \_\_init\_\_ method initializes the class. It takes three parameters: d\_model, seq\_len, and dropout. d\_model is the dimension of the embeddings, seq\_len is the maximum length of the sequences, and dropout is the dropout rate used for regularization. Inside this method, a dropout layer is created using nn.Dropout, and the positional encoding matrix pe is calculated and registered as a buffer.

The positional encoding matrix pe is calculated using a specific formula that uses sine and cosine functions of different frequencies. The formula is designed in such a way that the positional encoding for each position varies smoothly from -1 to 1, and the difference between the encodings of any two positions is small.

The forward method defines the forward pass of the module. It takes an input x which is expected to be a 3D tensor with shape [batch\_size, seq\_len, d\_model], representing the embedded input sequences. The method adds the positional encoding to the input embeddings (with the positional encoding adapted to the sequence length of x), applies dropout, and returns the result.

The output of the forward method is a 3D tensor with shape [batch\_size, seq\_len, d\_model], representing the input sequences with positional information added.

**Positional Encoding – Notebook**

This use case demonstrates how to use the PositionalEncoding class to add positional information to the input embeddings.

Here's a step-by-step explanation:

1. Import the necessary modules: sys (to modify the Python path), PyTorch, and the PositionalEncoding class.
2. Define the sequence length (seq\_len), the embedding dimension (d\_model), and the dropout rate (dropout). In this case, the sequence length is 50 (i.e., each sequence has 50 words), the embedding dimension is 300 (i.e., each word is represented as a 300-dimensional vector), and the dropout rate is 0.1.
3. Create an instance of the PositionalEncoding class, passing the embedding dimension, the sequence length, and the dropout rate as arguments.
4. Define a batch of 2 sequences, each sequence containing 50 words. The words are represented by their embeddings, which are 300-dimensional vectors. This is done using a PyTorch tensor filled with random numbers.
5. Pass the sequences through the PositionalEncoding instance. This adds positional information to the input embeddings and applies dropout.
6. Print the shape of the resulting tensor. The output shape is [2, 50, 300], which corresponds to [batch\_size, seq\_len, d\_model]. This means there are 2 sequences in the batch, each sequence has 50 words, and each word is represented as a 300-dimensional vector.

**Multi-Head Attention**

This class implements the Multi-Head Attention mechanism, a key component of the Transformer model.

Here's a step-by-step explanation:

1. \_\_init\_\_: This method initializes the class. It takes three parameters: d\_model (the feature length of a token), h (the number of attention heads), and dropout (the dropout rate). It creates four linear layers (w\_q, w\_k, w\_v, w\_o) and a dropout layer. The linear layers are used to transform the input sequences into queries, keys, and values, and to transform the output of the attention mechanism. The dropout layer is used to apply dropout to the attention scores.
2. attention: This static method calculates the attention scores and the output of the attention mechanism. It takes six parameters: query\_k, key\_k, value\_k (the queries, keys, and values), d\_k (the dimension of the key vectors), mask (an optional mask), and dropout (a dropout layer). It first calculates the attention scores by taking the dot product of the queries and keys, dividing by the square root of d\_k, and applying softmax. If a mask is provided, it is applied to the attention scores before softmax. Then, it multiplies the attention scores with the values to get the output of the attention mechanism.
3. forward: This method applies the multi-head attention mechanism to the input sequences. It takes four parameters: query, key, value (the input sequences), and mask (an optional mask). It first transforms the input sequences into queries, keys, and values using the linear layers. Then, it reshapes these tensors to have a separate dimension for the attention heads. After that, it calls the attention method to calculate the attention scores and the output of the attention mechanism. Finally, it concatenates the outputs of the different attention heads and transforms the result using the w\_o linear layer.

The output of the forward method is a tensor of the same shape as the input sequences, but the values have been updated to reflect the attention mechanism. This tensor can be used as input to the next layer of the Transformer model.

**Attention**

The attention function is a key part of the Transformer model architecture, specifically used in the self-attention mechanism. It calculates the attention scores for a given query, key, and value.

Here's a breakdown of the function:

1. The function takes five parameters: query\_k, key\_k, value\_k, d\_k, and optionally mask and dropout. query\_k, key\_k, and value\_k are the query, key, and value vectors, respectively. d\_k is the dimension of the key vectors. mask is an optional parameter used to prevent attention to certain positions. dropout is a dropout layer used for regularization.
2. The attention score is calculated by taking the dot product of the query and key (with the key transposed), and dividing by the square root of d\_k. This results in a tensor of shape [batch\_size, h, seq\_len, seq\_len].
3. If a mask is provided, the function applies the mask to the attention scores. This is done by replacing the scores at the masked positions with a very large negative number (-1e9), which becomes close to zero after applying the softmax function.
4. The softmax function is applied to the attention scores to convert them into probabilities. The probabilities are then passed through the dropout layer for regularization.
5. Finally, the function returns the weighted sum of the value vectors (calculated by taking the dot product of the attention scores and the value vectors), and the attention scores. The weighted sum represents the output of the attention mechanism, and the attention scores can be used for visualization or other purposes.

The output of the function is a pair of tensors with shapes [batch\_size, h, seq\_len, d\_k] and [batch\_size, h, seq\_len, seq\_len], respectively.

**Attention – Notebook**

**Usage of Attention without Mask**

This code is using the attention function from the MultiHeadAttention class to compute the attention scores and the output of the attention mechanism for a given batch of sequences. In this case, no mask is applied, so all positions can attend to all other positions.

Here's a step-by-step explanation:

1. Import the necessary modules: sys (to modify the Python path), PyTorch, and the MultiHeadAttention class.
2. Define the batch size, number of attention heads, sequence length, and dimension of the key/query vectors. In this case, there is 1 sequence in the batch, the sequence has 4 words, there is 1 attention head, and the dimension of the key/query vectors is 64.
3. Create an input tensor filled with random numbers. This tensor represents the input sequence.
4. Create the query, key, and value vectors by cloning the input tensor. These vectors are used as input to the attention mechanism.
5. Call the attention function, passing the query, key, and value vectors, the dimension of the key vectors, and a dropout layer as arguments. No mask is passed, so all positions can attend to all other positions.
6. Print the shape of the output tensor, the shape of the attention scores tensor, and the attention scores themselves. The output tensor is a weighted sum of the value vectors, and the attention scores represent how much each word in the sequence should attend to each other word in the sequence.

The output of the attention function is a pair of tensors: the output of the attention mechanism and the attention scores. The shapes of these tensors are printed to verify that they are as expected (the output tensor should have the same shape as the input tensor, and the attention scores tensor should have a shape of [batch\_size, h, seq\_len, seq\_len]).

## Example of Attention with Mask

This code is similar to the previous one, but it introduces a mask to the attention mechanism. The mask is used to prevent certain positions from attending to other positions. This is useful in certain scenarios, such as when you want to prevent future positions from being used in the prediction of the current position (causal masking), or when you want to ignore padding positions.

Here's a step-by-step explanation:

1. The first few steps are the same as before: import the necessary modules, define the batch size, number of attention heads, sequence length, and dimension of the key/query vectors, and create the input tensor and the query, key, and value vectors.
2. Create a causal mask. This is a square matrix of ones with zeros above the diagonal. This mask is used to prevent each position from attending to future positions. The triu function is used to create an upper triangular matrix, and the result is inverted to get a lower triangular matrix.
3. Create a padding mask. This is a vector that indicates which positions are padding (represented by zeros) and which are not (represented by ones). In this case, the last two positions are padding.
4. Combine the causal mask and the padding mask using the bitwise AND operator. This creates a mask that prevents each position from attending to future positions and padding positions.
5. Call the attention function, passing the query, key, and value vectors, the dimension of the key vectors, the mask, and a dropout layer as arguments.
6. Print the mask, the shape of the output tensor, the shape of the attention scores tensor, and the attention scores themselves.

The output of the attention function is a pair of tensors: the output of the attention mechanism and the attention scores. The shapes of these tensors are printed to verify that they are as expected. The mask is also printed for reference.

**Multi-Head Attention Notebook**

This code is using the MultiHeadAttention class to apply the multi-head attention mechanism to a batch of sequences. Here's a step-by-step explanation:

1. Import the necessary modules: sys (to modify the Python path), PyTorch, and the MultiHeadAttention class.
2. Define the dimensions of the model (d\_model), the number of attention heads (h), the dropout rate, the batch size, and the sequence length. In this case, the model has 512 dimensions, there are 8 attention heads, the dropout rate is 0.1, the batch size is 10, and each sequence has 20 words.
3. Create an instance of the MultiHeadAttention class, passing the dimensions of the model, the number of attention heads, and the dropout rate as arguments.
4. Create a random tensor to represent a batch of sequences. This tensor has a shape of [batch\_size, seq\_len, d\_model], which corresponds to a batch of 10 sequences, each with 20 words, and each word represented by a 512-dimensional vector.
5. Create the query, key, and value tensors by cloning the input tensor. These tensors are used as input to the multi-head attention mechanism.
6. Pass the query, key, and value tensors through the multi-head attention layer by calling the forward method of the MultiHeadAttention instance. This applies the multi-head attention mechanism to the input sequences and returns a new tensor that represents the output of the attention mechanism.
7. Print the shape of the output tensor. This tensor has the same shape as the input tensor ([batch\_size, seq\_len, d\_model]), but the values have been updated to reflect the attention mechanism. This tensor can be used as input to the next layer of the Transformer model.

**Feed Forward Layer**

This code defines a FeedForward layer, which is a component of the Transformer model. The FeedForward layer is a simple fully connected neural network that is applied to each position separately and identically. This includes two linear transformations with a ReLU activation in between.

Here's a step-by-step explanation:

1. The FeedForward class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, three layers are defined: linear1, linear2, and dropout. linear1 and linear2 are linear (fully connected) layers, and dropout is a dropout layer for regularization. The dimensions of the input and output of the linear layers (d\_model and d\_ff) are passed as arguments to the \_\_init\_\_ method.
3. The forward method is where the actual computation happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model].
4. The input tensor is first passed through linear1, which applies a linear transformation to every element of the input.
5. The output of linear1 is then passed through a ReLU activation function. The ReLU function is applied element-wise, and it replaces negative values with zeros.
6. The output of the ReLU function is then passed through the dropout layer, which randomly sets some elements to zero with a probability equal to dropout. This helps prevent overfitting.
7. The output of the dropout layer is then passed through linear2, which applies another linear transformation.
8. The result is a tensor of the same shape as the input tensor, but with the values transformed by two linear layers and a ReLU activation function. This tensor is returned as the output of the layer.

**Feed Forward Layer Notebook**

This code is using the FeedForward class to apply a feed-forward neural network to a batch of sequences. Here's a step-by-step explanation:

1. Import the necessary modules: sys (to modify the Python path), PyTorch, and the FeedForward class.
2. Define the dimensions of the model (d\_model), the dimensions of the feed-forward network (d\_ff), the dropout rate, the batch size, and the sequence length. In this case, each word is represented by a 7-dimensional vector, the feed-forward network has 2048 dimensions, the dropout rate is 0.1, the batch size is 1, and each sequence has 4 words.
3. Create an instance of the FeedForward class, passing the dimensions of the model, the dimensions of the feed-forward network, and the dropout rate as arguments.
4. Set a seed for PyTorch's random number generator to ensure that the same random numbers are generated every time the program is run.
5. Create a random tensor to represent a batch of sequences. This tensor has a shape of [batch\_size, seq\_len, d\_model], which corresponds to a batch of 1 sequence, with 4 words, and each word represented by a 7-dimensional vector.
6. Pass the tensor through the feed-forward network by calling the forward method of the FeedForward instance. This applies the feed-forward network to the input sequences and returns a new tensor that represents the output of the network.
7. Print the shape of the output tensor. This tensor has the same shape as the input tensor ([batch\_size, seq\_len, d\_model]), but the values have been transformed by the feed-forward network.
8. Print the output tensor. This tensor represents the output of the feed-forward network for each word in each sequence in the batch.

**Normalization Layer**

This code defines a Layer Normalization (LayerNorm) module. Layer normalization is a type of normalization technique like Batch Normalization, but instead of normalizing the features across the batch, it normalizes the features across the feature dimension (i.e., for each individual sample in the batch). This makes it batch size independent and can be used in a variety of contexts, including RNNs and Transformer models.

Here's a step-by-step explanation:

1. The LayerNorm class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, two parameters are defined: para\_mul and para\_bias. These are learnable parameters that will be optimized during training. They are initialized to ones and zeros, respectively. The eps (epsilon) value is a small constant used for numerical stability to avoid division by zero.
3. The forward method is where the actual normalization happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model].
4. The mean and standard deviation of the input tensor are computed along the last dimension (i.e., the feature dimension). The keepdim=True argument ensures that the output tensors have the same number of dimensions as the input tensor.
5. The input tensor is then normalized by subtracting the mean and dividing by the standard deviation. The eps value is added to the denominator to prevent division by zero.
6. The normalized tensor is then scaled by para\_mul and shifted by para\_bias. These are learnable parameters that allow the layer to learn an optimal scale and mean for the outputs.
7. The result is a tensor of the same shape as the input tensor, but with normalized values. This tensor is returned as the output of the layer.

**Normalization Layer Notebook**

This code is using the LayerNorm class to apply layer normalization to a batch of sequences. Here's a step-by-step explanation:

1. Import the necessary modules: sys (to modify the Python path), PyTorch, and the LayerNorm class.
2. Define the feature length (d\_model). In this case, each word is represented by a 7-dimensional vector.
3. Create an instance of the LayerNorm class, passing the feature length as an argument.
4. Set a seed for PyTorch's random number generator to ensure that the same random numbers are generated every time the program is run.
5. Create a random tensor to represent a batch of sequences. This tensor has a shape of [1, 4, 7], which corresponds to a batch of 1 sequence, with 4 words, and each word represented by a 7-dimensional vector.
6. Pass the tensor through the layer normalization layer by calling the forward method of the LayerNorm instance. This applies layer normalization to the input sequences and returns a new tensor that represents the normalized sequences.
7. Print the shape of the output tensor. This tensor has the same shape as the input tensor ([1, 4, 7]), but the values have been normalized.
8. Print the normalized sequences. These sequences have the same shape as the input sequences, but their values have been scaled and shifted to have a mean of 0 and a standard deviation of 1 (approximately, due to the learnable parameters and the epsilon value used for numerical stability).

**Residual Connection**

This code defines a ResidualConnection layer, which is a key component of the Transformer model. The ResidualConnection layer applies a sublayer to the input, and then adds the original input back to the result. This "shortcut" helps to mitigate the problem of vanishing gradients in deep networks, allowing the model to learn more effectively.

Here's a step-by-step explanation:

1. The ResidualConnection class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, a Layer Normalization (LayerNorm) and a Dropout layer are defined. The Layer Normalization normalizes the features across the feature dimension, and the Dropout layer randomly sets some elements to zero with a probability equal to dropout for regularization.
3. The forward method is where the actual computation happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model] and a sublayer function sublayer.
4. The input tensor is first normalized using the Layer Normalization.
5. The normalized tensor is then passed through the sublayer function. This could be any function that takes a tensor as input and returns a tensor of the same shape, such as a feed-forward network or a self-attention mechanism.
6. The output of the sublayer function is then passed through the Dropout layer.
7. The original input tensor x is added to the output of the Dropout layer. This is the "residual connection" that gives the layer its name.
8. The result is a tensor of the same shape as the input tensor, but with the values transformed by the sublayer and then added back to the original input. This tensor is returned as the output of the layer.

**Residual Connection Notebook**

This code is using the ResidualConnection class to apply a residual connection to a batch of sequences, with a feed-forward network as the sublayer. Here's a step-by-step explanation:

1. Import the necessary modules: sys (to modify the Python path), PyTorch, and the FeedForward, LayerNorm, and ResidualConnection classes.
2. Define the dimensions of the model (d\_model), the dimensions of the feed-forward network (d\_ff), the dropout rate, the batch size, and the sequence length. In this case, each word is represented by a 7-dimensional vector, the feed-forward network has 2048 dimensions, the dropout rate is 0.1, the batch size is 1, and each sequence has 4 words.
3. Create an instance of the ResidualConnection class, passing the dimensions of the model and the dropout rate as arguments.
4. Create an instance of the FeedForward class, passing the dimensions of the model, the dimensions of the feed-forward network, and the dropout rate as arguments.
5. Set a seed for PyTorch's random number generator to ensure that the same random numbers are generated every time the program is run.
6. Create a random tensor to represent a batch of sequences. This tensor has a shape of [batch\_size, seq\_len, d\_model], which corresponds to a batch of 1 sequence, with 4 words, and each word represented by a 7-dimensional vector.
7. Pass the tensor through the residual connection by calling the forward method of the ResidualConnection instance, with the FeedForward instance as the sublayer. This applies the feed-forward network to the normalized input sequences, applies dropout, and then adds the original input back to the result.
8. Print the original input tensor, the shape of the output tensor, and the output tensor. The output tensor has the same shape as the input tensor ([batch\_size, seq\_len, d\_model]), but the values have been transformed by the feed-forward network and then added back to the original input.

**Projection Layer**

This code defines a Projection layer, which is a component of the Transformer model. The Projection layer is a simple fully connected neural network that is used to transform the output of the Transformer's decoder into prediction scores for each possible output token in the vocabulary.

Here's a step-by-step explanation:

1. The Projection class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, a linear layer named projection is defined. This layer will map from the dimensionality of the model (d\_model) to the size of the output vocabulary (vocab\_size).
3. The forward method is where the actual computation happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model].
4. The input tensor is passed through the projection layer, which applies a linear transformation to every element of the input.
5. The result is a tensor of shape [batch\_size, seq\_len, vocab\_size]. This tensor represents the prediction scores for each possible output token in the vocabulary, for each position in each sequence in the batch.
6. This tensor is returned as the output of the layer. In a full Transformer model, these scores would typically be passed through a softmax function to produce a probability distribution over the vocabulary for each position in each sequence.

**Projection Notebook**

This code is using the Projection class to apply a projection to a batch of sequences. Here's a step-by-step explanation:

1. Import the necessary modules: sys (to modify the Python path), PyTorch, and the Projection class.
2. Define the dimensions of the model (d\_model), the size of the output vocabulary (vocab\_size), the batch size, and the sequence length. In this case, each word is represented by a 7-dimensional vector, the output vocabulary has 15 tokens, the batch size is 1, and each sequence has 4 words.
3. Create an instance of the Projection class, passing the dimensions of the model and the size of the output vocabulary as arguments.
4. Set a seed for PyTorch's random number generator to ensure that the same random numbers are generated every time the program is run.
5. Create a random tensor to represent a batch of sequences. This tensor has a shape of [batch\_size, seq\_len, d\_model], which corresponds to a batch of 1 sequence, with 4 words, and each word represented by a 7-dimensional vector.
6. Pass the tensor through the projection by calling the forward method of the Projection instance. This applies the projection to the input sequences and returns a new tensor that represents the prediction scores for each possible output token in the vocabulary.
7. Print the original input tensor, the shape of the output tensor, and the output tensor. The output tensor has a shape of [batch\_size, seq\_len, vocab\_size], which means it contains prediction scores for 15 possible output tokens for each of the 4 words in each of the 1 sequence in the batch.

---

Now I am going to implement the class and function to create transformers architecture, in which I will show you Encoder Layer, Encoder, Decoder Layer, Decoder, Transformer class, and finally the function create transformer model that combines all the building blocks into the final transformer model. Let’s deep dive in.

---

**Encoder Layer**

This code defines an EncoderLayer class, which is a component of the Transformer model. Each EncoderLayer consists of two main parts: a multi-head self-attention mechanism and a position-wise feed-forward network. These parts are connected by residual connections and followed by layer normalization.

Here's a step-by-step explanation:

1. The EncoderLayer class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, the multi-head self-attention mechanism, the feed-forward network, and the residual connections are defined. The residual connections are a list of two ResidualConnection modules, which are used to add the input of each part to its output.
3. The forward method is where the actual computation happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model] and an optional source mask mask\_scr.
4. The input tensor is first passed through the first residual connection and the self-attention mechanism. The self-attention mechanism allows the model to focus on different parts of the input sequence for each output position. The lambda function is used to ensure that the same input is used for the query, key, and value in the self-attention mechanism.
5. The output of the self-attention mechanism is then passed through the second residual connection and the feed-forward network. The feed-forward network applies the same transformation to each position in the sequence independently.
6. The output of the feed-forward network is returned as the output of the layer. This tensor has the same shape as the input tensor and can be passed to the next EncoderLayer in the sequence.

In a full Transformer model, several EncoderLayers are stacked on top of each other to form the encoder. Each EncoderLayer operates independently on the input, allowing the model to learn complex patterns in the data.

**Encoder**

This code defines an Encoder class, which is a component of the Transformer model. The Encoder consists of a stack of identical layers, each of which is an instance of the EncoderLayer class.

Here's a step-by-step explanation:

1. The Encoder class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, a list of EncoderLayer instances is created. The number of layers is specified by the num\_layers argument. A layer normalization module (LayerNorm) is also created. This is used to normalize the output of the last layer.
3. The forward method is where the actual computation happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model] and an optional source mask mask\_scr.
4. The input tensor is passed through each EncoderLayer in the list. Each layer applies a multi-head self-attention mechanism and a feed-forward network to the input, and adds the result to the original input (residual connection). The output of each layer is used as the input to the next layer.
5. After all layers have processed the input, the output of the last layer is normalized using the layer normalization module. This helps to stabilize the learning process and reduces the training time.
6. The normalized output is returned as the output of the encoder. This tensor has the same shape as the input tensor and can be passed to the decoder of the Transformer model.

In a full Transformer model, the Encoder processes the source sequences and produces a set of high-level features that represent the content of the sequences. These features are then used by the Decoder to generate the target sequences.

**Encoder Notebook**

This code is using an Encoder to process a batch of sequences. Here's a step-by-step explanation:

1. Import the necessary modules and classes.
2. Define the configuration parameters: the dimensions of the model (d\_model), the dimensions of the feed-forward network (d\_ff), the number of heads in the multi-head attention mechanism (h), the batch size, the sequence length, the dropout ratio, and the number of layers in the encoder (num\_layers).
3. Create instances of the MultiHeadAttention and FeedForward classes. These are the main components of the EncoderLayer.
4. Create an instance of the EncoderLayer class, passing the dimensions of the model, the multi-head attention mechanism, the feed-forward network, and the dropout ratio as arguments.
5. Create an instance of the Encoder class, passing the dimensions of the model, the EncoderLayer instance, and the number of layers as arguments. This creates an encoder with num\_layers identical layers.
6. Set a seed for PyTorch's random number generator to ensure that the same random numbers are generated every time the program is run.
7. Create a random tensor to represent a batch of sequences. This tensor has a shape of [batch\_size, seq\_len, d\_model], which corresponds to a batch of 1 sequence, with 4 words, and each word represented by an 8-dimensional vector.
8. Pass the tensor through the encoder by calling the forward method of the Encoder instance. This applies the multi-head attention mechanism and the feed-forward network to the input sequences and returns a new tensor that represents the output of the encoder.
9. Print the original input tensor, the shape of the output tensor, and the output tensor. The output tensor has the same shape as the input tensor, which means it can be passed to the decoder of the Transformer model.

**---**

**Decoder Layer**

This code defines a DecoderLayer class, which is a component of the Transformer model. Each DecoderLayer consists of three main parts: a self-attention mechanism, an encoder-decoder attention mechanism, and a position-wise feed-forward network. These parts are connected by residual connections and followed by layer normalization.

Here's a step-by-step explanation:

1. The DecoderLayer class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, the self-attention mechanism, the encoder-decoder attention mechanism, the feed-forward network, and the residual connections are defined. The residual connections are a list of three ResidualConnection modules, which are used to add the input of each part to its output.
3. The forward method is where the actual computation happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model], an encoder\_output tensor of the same shape, and optional source and target masks mask\_src and mask\_tgt.
4. The input tensor is first passed through the first residual connection and the self-attention mechanism. The self-attention mechanism allows the model to focus on different parts of the input sequence for each output position. The lambda function is used to ensure that the same input is used for the query, key, and value in the self-attention mechanism. The target mask is used to prevent the decoder from looking at future tokens.
5. The output of the self-attention mechanism is then passed through the second residual connection and the encoder-decoder attention mechanism. This mechanism allows the decoder to focus on different parts of the encoder output for each output position. The source mask is used to prevent the decoder from looking at padding tokens.
6. The output of the encoder-decoder attention mechanism is then passed through the third residual connection and the feed-forward network. The feed-forward network applies the same transformation to each position in the sequence independently.
7. The output of the feed-forward network is returned as the output of the layer. This tensor has the same shape as the input tensor and can be passed to the next DecoderLayer in the sequence.

In a full Transformer model, several DecoderLayers are stacked on top of each other to form the decoder. Each DecoderLayer operates independently on the input, allowing the model to learn complex patterns in the data.

**Decoder**

This code defines a Decoder class, which is a component of the Transformer model. The Decoder consists of a stack of identical layers, each of which is a DecoderLayer, and a final layer normalization.

Here's a step-by-step explanation:

1. The Decoder class is a subclass of nn.Module, which means it's a custom PyTorch module.
2. In the \_\_init\_\_ method, the layers of the decoder and the final layer normalization are defined. The layers are a list of DecoderLayer instances, and the layer normalization is an instance of the LayerNorm class.
3. The forward method is where the actual computation happens. It takes an input tensor x of shape [batch\_size, seq\_len, d\_model], an encoder\_output tensor of the same shape, and optional source and target masks mask\_src and mask\_tgt.
4. The input tensor is passed through each layer in the decoder in turn. Each layer applies the self-attention mechanism, the encoder-decoder attention mechanism, and the feed-forward network to the input and encoder output, and returns a new tensor that represents the output of the layer.
5. After all layers have processed the input, the final layer normalization is applied. This ensures that the output has a mean of 0 and a standard deviation of 1, which can help to stabilize the learning process.
6. The output of the layer normalization is returned as the output of the decoder. This tensor has the same shape as the input tensor and can be passed to the next component of the Transformer model.

In a full Transformer model, the Decoder takes the output of the Encoder and generates a sequence of output tokens. Each DecoderLayer in the Decoder operates independently on the input, allowing the model to learn complex patterns in the data.

**Decoder Notebook**

This code is using a Decoder to process a batch of sequences. Here's a step-by-step explanation:

1. Import the necessary modules and classes.
2. Define the configuration parameters: the dimensions of the model (d\_model), the dimensions of the feed-forward network (d\_ff), the number of heads in the multi-head attention mechanism (h), the batch size, the sequence length, the dropout ratio, and the number of layers in the decoder (num\_layers).
3. Create instances of the MultiHeadAttention and FeedForward classes. These are the main components of the DecoderLayer.
4. Create an instance of the DecoderLayer class, passing the dimensions of the model, the self-attention mechanism, the encoder-decoder attention mechanism, the feed-forward network, and the dropout ratio as arguments.
5. Create an instance of the Decoder class, passing the dimensions of the model, the DecoderLayer instance, and the number of layers as arguments.
6. Set a seed for PyTorch's random number generator to ensure that the same random numbers are generated every time the program is run.
7. Create a random tensor to represent a batch of sequences. This tensor has a shape of [batch\_size, seq\_len, d\_model], which corresponds to a batch of 1 sequence, with 4 words, and each word represented by an 8-dimensional vector.
8. Create another random tensor to represent the output of the encoder. This tensor has the same shape as the input tensor.
9. Pass the input tensor and the encoder output through the decoder by calling the forward method of the Decoder instance. This applies the decoder layers and the final layer normalization to the input and encoder output, and returns a new tensor that represents the output of the decoder.
10. Print the original input tensor, the encoder output tensor, the shape of the output tensor, and the output tensor. The output tensor has the same shape as the input tensor, which means it can be passed to the next component of the Transformer model.

**---**

**Transformer Model Class**

The Transformer class is a PyTorch module that represents a Transformer model for sequence-to-sequence tasks like machine translation. Here's a step-by-step explanation:

1. Initialization (\_\_init\_\_): The class is initialized with an encoder, a decoder, source and target input embeddings, source and target positional encodings, and a projection layer. These are all stored as attributes of the class.
2. Encoding (encode): This method takes a source sequence and an optional source mask as input. It first applies the source input embedding and the source positional encoding to the source sequence, and then passes the result through the encoder. The output of the encoder is returned.
3. Decoding (decode): This method takes a target sequence, the output of the encoder, and optional source and target masks as input. It first applies the target input embedding and the target positional encoding to the target sequence, and then passes the result, along with the encoder output and the masks, through the decoder. The output of the decoder is returned.
4. Projection (project): This method takes a sequence (typically the output of the decoder) as input and passes it through the projection layer. The output of the projection layer, which has the same length as the input sequence but a depth equal to the target vocabulary size, is returned.

This class is typically used to create a Transformer model, which can then be trained on a sequence-to-sequence task. The encode, decode, and project methods correspond to the main steps of the Transformer model: encoding the source sequence, decoding the target sequence, and projecting the decoder output to the target vocabulary size.

**Create Transformer model**

This function create\_transformer\_model is used to create a Transformer model for a machine translation task. Here's a step-by-step explanation:

1. **Extract configuration parameters**: The function first extracts the necessary parameters from the configuration, such as the dimension of the model (d\_model), the number of layers (num\_layers), the number of attention heads (h), the dimension of the feed-forward network (d\_ff), the dropout rate (dropout), and the sequence length (seq\_len).
2. **Initialize embedding layers**: It initializes the source and target embedding layers using the InputEmbedding class. These layers convert the input tokens into vectors of dimension d\_model.
3. **Initialize positional encoding layers**: It initializes the source and target positional encoding layers using the PositionalEncoding class. These layers add positional information to the input embeddings.
4. **Initialize encoder**: It initializes the self-attention and feed-forward layers for the encoder using the MultiHeadAttention and FeedForward classes, respectively. It then creates an encoder layer using the EncoderLayer class and repeats this layer num\_layers times to create the encoder.
5. **Initialize decoder**: It initializes the self-attention, encoder-decoder attention, and feed-forward layers for the decoder. It then creates a decoder layer using the DecoderLayer class and repeats this layer num\_layers times to create the decoder.
6. **Initialize projection layer**: It initializes the projection layer using the Projection class. This layer converts the output of the decoder into logits for each possible output token.
7. **Initialize transformer model**: It creates the Transformer model using the Transformer class, passing in the encoder, decoder, embedding layers, positional encoding layers, and projection layer.
8. **Initialize model parameters**: It initializes the parameters of the model using the Xavier uniform initialization. This is a common initialization method for neural networks.
9. **Return transformer model**: Finally, the function returns the created Transformer model. This model can be used to train and evaluate a machine translation task.

**Put it all together - Playground Transformer**

This code is using a Transformer model to encode a batch of source sequences, decode a batch of target sequences, project the decoder output to the vocabulary size, and then decode the projected output into words. Here's a step-by-step explanation:

1. Import the necessary modules and functions.
2. Load the configuration file, dataset, and tokenizers for the source and target languages.
3. Define the vocabulary sizes for the source and target languages, the sequence length, and the batch size.
4. Create the Transformer model using the create\_tranformer\_model function.
5. Set a seed for PyTorch's random number generator to ensure that the same random numbers are generated every time the program is run.
6. Create a random tensor to represent a batch of source sequences and pass it through the encoder of the Transformer model.
7. Create another random tensor to represent a batch of target sequences and pass it along with the encoder output through the decoder of the Transformer model.
8. Pass the decoder output through the projection layer of the Transformer model. This transforms the decoder output into prediction scores for each possible output token in the vocabulary.
9. Find the token with the highest prediction score in each position of each sequence. This is done by calling torch.max on the projected output and taking the indices of the maximum values. The result is a tensor of shape [batch\_size, seq\_len], where each element is the index of the most probable token.
10. Convert the tensor of predicted tokens into a NumPy array. This is done by first removing any extra dimensions with squeeze, then detaching the tensor from the computation graph with detach, and finally converting it to a NumPy array with cpu().numpy().
11. Find the index of the first end-of-sentence (EOS) token in the predicted tokens. This is done by iterating over the predicted tokens and breaking the loop when the EOS token (which has an index of 3) is found.
12. Slice the predicted tokens up to the EOS token. This gives the predicted token sequence.
13. Decode the predicted token sequence into words using the target tokenizer. This converts the indices back into words.
14. Print the predicted sequence of words. This is the output of the Transformer model for the given input sequences.

---

Now I’m going to present to you the functions for launching training process, which includes the train, evaluation and inference during training function. Let’s deep dive in.

---

**Softmax**

Another important built-in function that we meet when implementing transformer is Softmax function.

Let’s see a concreate usage example of this function and what is its expected result.

The softmax function is a function used in machine learning, especially in the field of neural networks, to map a vector of arbitrary real-valued scores to a vector of probabilities.

The softmax function takes an N-dimensional vector of real numbers and transforms it into a vector of real number in range (0, 1) which add up to 1. This property makes it suitable to interpret the outputs of the model as probabilities.

The softmax function S for a vector Z of length K is defined as follows:

S(Z)\_i = exp(Z\_i) / sum(exp(Z\_j) for j in range(K))

Here, exp is the exponential function, Z\_i is the i-th component of the input vector Z, and sum(exp(Z\_j) for j in range(K)) is the sum of the exponential of all components of Z.

In other words, the softmax function exponentiates its input (making all components positive) and then normalizes it (making the sum of all components equal to 1).

This Python script applies the softmax function to an input tensor using PyTorch.

Here's a step-by-step explanation:

1. The script first imports the necessary libraries: PyTorch and NumPy.
2. It then creates an input tensor with three elements: 0.123, 0.264, and -np.inf. The -np.inf represents negative infinity.
3. The script applies the softmax function to the input tensor along dimension 0 (since the tensor is 1-dimensional, this is the only dimension). The softmax function is often used in the final layer of a neural network model to represent a probability distribution over n different outcomes. It transforms each element of the tensor into a value between 0 and 1, such that the sum of all elements in the tensor equals 1.
4. Finally, the script prints the input tensor and the output tensor. The output tensor should contain the softmax-transformed values of the input tensor. Note that the softmax of -np.inf is 0, as the softmax function effectively "squashes" very negative numbers to 0.

---

**Loss function – Cross Entropy Loss**

Before we deep dive into the training process, I would like to explain the key element of training process, that is calculating loss function. In transformer, they use CrossEntropyLoss to calculate the loss. Let’s see how to Implement this step and what is the expected result.

nn.CrossEntropyLoss is a PyTorch class that computes the cross-entropy loss. It's often used in multi-class classification problems

In this example, the model predicts the probabilities of 3 classes for 2 samples. The true classes for the samples are 2 and 0, respectively. The nn.CrossEntropyLoss function computes the cross-entropy loss between the predictions and the true classes.

Note that the target tensor in nn.CrossEntropyLoss is expected to contain the class indices, not the one-hot encoded vectors. Also, the output tensor is expected to contain raw, unnormalized scores for each class (also known as logits), not probabilities. The nn.CrossEntropyLoss function applies the softmax function to the output tensor internally to convert these scores into probabilities and then computes the negative log likelihood loss.

This will give you the same result as nn.CrossEntropyLoss. Note that we take the mean of the negative log likelihood loss over all samples in the batch, because nn.CrossEntropyLoss by default returns the mean loss. If you want the sum instead, you can replace .mean() with .sum().

This Python script calculates the cross-entropy loss between the output of a model and the target values.

Here's a step-by-step explanation:

1. The script first sets the size of the target vocabulary, the batch size, and the sequence length. It also sets the seed for PyTorch's random number generator to ensure reproducible results.
2. It then creates example tensors for the model's output (projection\_output) and the target values (decoder\_target). The projection\_output tensor is created with random values and has a shape of (batch\_size, seq\_len, vocab\_size\_tgt). The decoder\_target tensor is created with specific values and has a shape of (batch\_size, seq\_len).
3. The script assumes that the ID of the padding token is 0.
4. It initializes the cross-entropy loss function, specifying that the padding token should be ignored (ignore\_index=pad\_id) and that label smoothing should be applied (label\_smoothing=0.1). Label smoothing is a regularization technique that prevents the model from becoming too confident about its predictions.
5. The script reshapes the projection\_output and decoder\_target tensors to be 1-dimensional using the view method. This is necessary because the cross-entropy loss function expects the input and target tensors to be 1-dimensional.
6. It calculates the loss between the model's output and the target values using the cross-entropy loss function.
7. Finally, the script prints the reshaped output and target tensors and the calculated loss.

---

**Train**

The train function is responsible for training a Transformer model for a machine translation task. Here's a step-by-step explanation:

1. **Device Assignment**: It assigns the device (GPU if available, else CPU) for computation.
2. **Tensorboard Writer**: It initiates a Tensorboard writer for logging training metrics.
3. **Model Directory**: It checks if a directory for saving model weights exists, if not, it creates one.
4. **Data Preprocessing**: It calls the preprocessing\_data function to load and preprocess the dataset.
5. **Vocabulary Size**: It gets the vocabulary size for the source and target languages.
6. **Model Creation**: It creates the Transformer model using the create\_transformer\_model function.
7. **Optimizer Initialization**: It initializes the Adam optimizer.
8. **Checkpoint Loading**: If a checkpoint exists, it loads the model and optimizer states from the checkpoint, and sets the initial epoch and global step accordingly.
9. **Loss Function**: It defines the CrossEntropyLoss function, ignoring the padding token and applying label smoothing.
10. **Training Loop**: It runs a training loop for a specified number of epochs. For each batch in each epoch, it:
    * Moves the inputs and targets to the device
    * Passes the inputs through the model to get the output
    * Calculates the loss between the output and the target
    * Logs the loss to Tensorboard
    * Performs backpropagation and updates the model parameters
    * Increments the global step
11. **Validation**: At the end of each epoch, it runs a validation step using the evaluation\_step function.
12. **Checkpoint Saving**: It saves a checkpoint of the model and optimizer states, the current epoch, and the global step.

This function is typically called to start the training process. It handles all aspects of training, including data loading, model creation, training loop, validation, and checkpointing.

**Evaluation during training**

The evaluation\_step function is used to evaluate a trained Transformer model on a validation dataset. Here's a step-by-step explanation:

1. **Model Evaluation Mode**: It sets the model to evaluation mode.
2. **Token IDs**: It gets the IDs of the start-of-sequence (SOS) and end-of-sequence (EOS) tokens.
3. **Text Lists**: It initializes lists to store the source texts, target texts, and predicted texts.
4. **Evaluation Loop**: It runs a loop over the validation dataloader. For each batch in the dataloader, it:
   * Moves the encoder input and mask to the device.
   * Asserts that the batch size is 1.
   * Calls the infer\_training function to generate a sequence of tokens from the model.
   * Decodes the source text, target text, and predicted text.
   * Appends the texts to the respective lists.
   * Prints the source text, target text, and predicted text.
   * Breaks the loop if the number of evaluated samples reaches the specified limit.
5. **Metrics Calculation**: If logs are enabled, it calculates the BLEU score, word error rate (WER), and character error rate (CER) between the predicted texts and target texts, and logs these metrics.

This function is typically used in the validation stage of a machine translation task, where the goal is to evaluate the performance of the trained model on unseen data. The evaluation metrics provide a quantitative measure of the model's performance.

---

**Inference during training**

This Python function, infer\_training, performs inference using a provided Transformer model. Here's a step-by-step explanation:

1. The function takes several arguments: a Transformer model, an input tensor for the encoder, a mask for the encoder input, the IDs of the start and end of sequence tokens, the length of the sequence, and the device to perform computations on.
2. It first encodes the input using the model's encode method, which returns the encoder output.
3. It then initializes the decoder input with the start of sequence token. The decoder input is a tensor of shape (1, 1).
4. The function enters a loop that continues until the sequence length of the decoder input equals the sequence length of the encoder input.
5. Inside the loop, it creates a causal mask for the decoder input. This mask is used in the Transformer model to prevent the decoder from "seeing" future tokens in the input.
6. It decodes the encoder output using the model's decode method, which returns the decoder output.
7. It selects the last token from the sequence length dimension of the decoder output.
8. It projects this token's output to the target vocabulary size using the model's project method.
9. It finds the token with the maximum value in the projected output, which is the predicted token.
10. It appends this predicted token to the decoder input.
11. If the predicted token is the end of sequence token, it breaks the loop.
12. Finally, it returns the decoder input, removing the batch dimension.

This function is used during the training of the Transformer model to perform inference on the training data. It generates a sequence of tokens from the model, starting with the start of sequence token and ending when the model predicts the end of sequence token or when the sequence length reaches the specified limit.

---

**Training in action**

Well, now we have built the whole transformer package. It’s time to call the training engine in action.

We need to write this small code snippet to do it.

This Python script is used to train a Transformer model. Here's a step-by-step explanation:

1. It first imports the necessary functions: train from the transformer.engine module and load\_config from the transformer.utils module.
2. The script then checks if it is being run as a standalone program (as opposed to being imported as a module) with the if \_\_name\_\_ == "\_\_main\_\_": statement.
3. Inside this conditional block, it sets the path of the configuration file to "config.json".
4. It loads the configuration from this file using the load\_config function. This function presumably reads the JSON file and returns a dictionary containing the configuration parameters.
5. Finally, it calls the train function with the loaded configuration as an argument. This function presumably trains the Transformer model according to the specified configuration.

In summary, this script is a driver script for training a Transformer model. It reads the training configuration from a JSON file and then calls a function to perform the training.

---

After that, we call the train.py on the terminal to launch the training.

The training process starts with checking the available device for training.

Then if there exists a training checkpoint since the last training, if so, it will load the weight from the last checkpoint and continue to train for a new checkpoint. If it does not exist the checkpoint, it will start training from the first epoch.

After one epoch is trained successfully, it will print several inference examples based on the trained model until that epoch. Then, the checkpoint model will saved in the model folder

When we finish training over all epochs, we will obtain all the checkpoint model like this.

At this point, we are ready to start inferencing and testing the latest trained model.

Let’s continue to this excited part.

---

Intro to Inference

Well, now I am going to show you how we can do inference on the trained model.

Given the task of machine translation in transformer, the inference process

* Firstly load the latest trained model, and the necessary configuration setting, like target tokenizers, etc
* Then ot takes a sentence in one language, in this case a English sentence, as input.
* Finally, it output a translated sentence in French.

Let’s deep dive into how we can obtain the inference process.

---

To do the inference process, we need to implement the inference function in the engine file

**Model Inference**

This Python function, inference, translates a source text using a Transformer model. Here's a step-by-step explanation:

1. The function takes several arguments: the source text to translate, a Transformer model, tokenizers for the source and target languages, the maximum sequence length, and the device to perform computations on.
2. It first sets the model to evaluation mode with model.eval(). This is necessary because the model behaves differently during training and evaluation.
3. It then disables gradient calculation with torch.no\_grad(). This is done because gradients are not needed during inference, and disabling their calculation saves memory.
4. It tokenizes the source text using the source language tokenizer's encode method.
5. It calculates the number of padding tokens needed for the encoder input. This is done by subtracting the length of the tokenized source text and 2 (for the start and end of sequence tokens) from the maximum sequence length.
6. It gets the IDs of the start of sequence, end of sequence, and padding tokens using the source language tokenizer's token\_to\_id method.
7. It creates the encoder input by concatenating the start of sequence token, the tokenized source text, the end of sequence token, and the necessary number of padding tokens.
8. It creates the encoder mask, which is used in the Transformer model to ignore the padding tokens during the attention calculation.
9. It encodes the input using the model's encode method, which returns the encoder output.
10. It initializes the decoder input with the start of sequence token.
11. It enters a loop that continues until the sequence length of the decoder input equals the maximum sequence length or the end of sequence token is predicted.
12. Inside the loop, it creates a causal mask for the decoder input. This mask is used in the Transformer model to prevent the decoder from "seeing" future tokens in the input.
13. It decodes the encoder output using the model's decode method, which returns the decoder output.
14. It selects the last token from the sequence length dimension of the decoder output.
15. It projects this token's output to the target vocabulary size using the model's project method.
16. It finds the token with the maximum value in the projected output, which is the predicted token.
17. It appends this predicted token to the decoder input.
18. If the predicted token is the end of sequence token, it breaks the loop.
19. Finally, it decodes the decoder input using the target language tokenizer's decode method to get the translated text, and returns this text.

This function is used during the inference phase of the Transformer model to translate a source text to the target language. It generates a sequence of tokens from the model, starting with the start of sequence token and ending when the model predicts the end of sequence token or when the sequence length reaches the specified limit.

**Playground inference**

This Python script translates a source text using a pre-trained Transformer model. Here's a step-by-step explanation:

1. It first imports the necessary modules and functions.
2. It then assigns the device for computation. If a GPU is available, it uses that; otherwise, it uses the CPU.
3. It loads the configuration from a JSON file using the load\_config function.
4. It gets the dataset using the get\_dataset function and the configuration.
5. It gets the tokenizers for the source and target languages using the get\_tokenizer function, the configuration, the dataset, and the respective language codes from the configuration.
6. It gets the vocabulary sizes for the source and target languages using the get\_vocab\_size method of the respective tokenizers.
7. It gets the maximum sequence length from the configuration.
8. It creates the Transformer model using the create\_tranformer\_model function, the configuration, and the vocabulary sizes. It moves the model to the assigned device.
9. It gets the path of the latest checkpoint using the get\_checkpoint\_path function and the configuration.
10. It loads the state of the model from the checkpoint using torch.load.
11. It assigns the loaded state to the Transformer model using the load\_state\_dict method.
12. It sets the source text to translate.
13. It calls the inference function with the source text, the model, the tokenizers, the maximum sequence length, and the device. This function translates the source text using the model and returns the translated text.
14. Finally, it prints the source text and the translated text.

In summary, this script loads a pre-trained Transformer model and uses it to translate a source text. It demonstrates how to use a Transformer model for inference.

---

Well, thank you so much for bearing with me until the end of this long journey. I hope you now have a good understanding of how to implement the Transformer model in practice using Python and the PyTorch framework.

If there are any points you don't understand, please let me know in the comments, and I will try to answer them. Or if you find any mistakes, please let me know so I can correct them and improve for next time.

Finally, if you like my video, you can encourage me by subscribing to my channel, liking, and sharing my video with your friends or on your social networks. You can also follow on Github and star my repository if you find it helpful. That helps others find it easily.

That will motivate me a lot for the upcoming videos. Let's learn and grow together. Thank you very much!