**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.

**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**

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.

**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.

**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).

**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.

**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.

**Testing model: encoder, decoder, tokenizing with random source and target text**

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.