```
START, torch.Size([1])
embeddings.weight, torch.Size([50264, 1024])
seq2seq_encoder.embeddings.weight, torch.Size([50264, 1024])
seq2seq_encoder.position_embeddings.weight, torch.Size([1024, 1024])
seq2seq_encoder.norm_embeddings.weight, torch.Size([1024])
seq2seq_encoder.norm_embeddings.bias, torch.Size([1024])
seq2seq_encoder.layers.0.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.0.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.0.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.0.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.0.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.0.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.0.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.0.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.0.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.0.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.0.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.0.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.0.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.0.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.0.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.0.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.1.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.1.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.1.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.1.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.1.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.1.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.1.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.1.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.1.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.1.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.1.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.1.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.1.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.1.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.1.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.1.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.2.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.2.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.2.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.2.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.2.attention.v_lin.weight, torch.Size([1024,
1024])
```

```
seq2seq_encoder.layers.2.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.2.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.2.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.2.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.2.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.2.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.2.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.2.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.2.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.2.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.2.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.3.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.3.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.3.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.3.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.3.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.3.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.3.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.3.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.3.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.3.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.3.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.3.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.3.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.3.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.3.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.3.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.4.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.4.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.4.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.4.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.4.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.4.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.4.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.4.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.4.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.4.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.4.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.4.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.4.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.4.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.4.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.4.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.5.attention.q_lin.weight, torch.Size([1024,
1024])
```

```
seq2seq_encoder.layers.5.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.5.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.5.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.5.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.5.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.5.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.5.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.5.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.5.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.5.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.5.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.5.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.5.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.5.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.5.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.6.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.6.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.6.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.6.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.6.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.6.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.6.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.6.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.6.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.6.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.6.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.6.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.6.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.6.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.6.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.6.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.7.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.7.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.7.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.7.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.7.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.7.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.7.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.7.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.7.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.7.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.7.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.7.ffn.lin1.bias, torch.Size([4096])
```

```
seq2seq_encoder.layers.7.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.7.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.7.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.7.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.8.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.8.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.8.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.8.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.8.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.8.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.8.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.8.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.8.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.8.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.8.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.8.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.8.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.8.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.8.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.8.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.9.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.9.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.9.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.9.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.9.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.9.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.9.attention.out_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.9.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.9.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.9.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.9.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.9.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.9.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.9.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.9.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.9.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.10.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.10.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.10.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.10.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.10.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.10.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.10.attention.out_lin.weight,
```

```
torch.Size([1024, 1024])
seq2seq_encoder.layers.10.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.10.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.10.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.10.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.10.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.10.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.10.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.10.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.10.norm2.bias, torch.Size([1024])
seq2seq_encoder.layers.11.attention.q_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.11.attention.q_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.11.attention.k_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.11.attention.k_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.11.attention.v_lin.weight, torch.Size([1024,
1024])
seq2seq_encoder.layers.11.attention.v_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.11.attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_encoder.layers.11.attention.out_lin.bias, torch.Size([1024])
seq2seq_encoder.layers.11.norm1.weight, torch.Size([1024])
seq2seq_encoder.layers.11.norm1.bias, torch.Size([1024])
seq2seq_encoder.layers.11.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_encoder.layers.11.ffn.lin1.bias, torch.Size([4096])
seq2seq_encoder.layers.11.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_encoder.layers.11.ffn.lin2.bias, torch.Size([1024])
seq2seq_encoder.layers.11.norm2.weight, torch.Size([1024])
seq2seq_encoder.layers.11.norm2.bias, torch.Size([1024])
retriever.query_encoder.embeddings.weight, torch.Size([30522, 768])
retriever.query_encoder.position_embeddings.weight, torch.Size([512,
768])
retriever.query_encoder.norm_embeddings.weight, torch.Size([768])
retriever.query_encoder.norm_embeddings.bias, torch.Size([768])
retriever.query_encoder.segment_embeddings.weight, torch.Size([2,
768])
retriever.query_encoder.layers.0.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.0.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.0.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.0.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.0.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.0.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.0.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.0.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.0.norm1.weight, torch.Size([768])
```

```
retriever.query_encoder.layers.0.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.0.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.0.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.0.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.0.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.0.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.0.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.1.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.1.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.1.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.1.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.1.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.1.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.1.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.1.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.1.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.1.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.1.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.1.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.1.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.1.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.1.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.1.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.2.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.2.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.2.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.2.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.2.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.2.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.2.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.2.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.2.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.2.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.2.ffn.lin1.weight, torch.Size([3072,
```

768])
retriever.query_encoder.layers.2.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.2.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.2.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.2.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.2.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.3.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.3.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.3.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.3.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.3.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.3.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.3.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.3.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.3.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.3.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.3.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.3.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.3.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.3.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.3.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.3.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.4.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.4.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.4.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.4.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.4.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.4.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.4.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.4.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.4.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.4.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.4.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.4.ffn.lin1.bias, torch.Size([3072])

```
retriever.query_encoder.layers.4.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.4.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.4.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.4.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.5.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.5.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.5.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.5.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.5.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.5.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.5.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.5.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.5.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.5.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.5.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.5.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.5.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.5.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.5.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.5.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.6.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.6.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.6.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.6.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.6.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.6.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.6.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.6.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.6.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.6.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.6.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.6.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.6.ffn.lin2.weight, torch.Size([768,
3072])
```

```
retriever.query_encoder.layers.6.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.6.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.6.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.7.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.7.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.7.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.7.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.7.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.7.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.7.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.7.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.7.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.7.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.7.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.7.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.7.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.7.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.7.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.7.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.8.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.8.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.8.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.8.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.8.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.8.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.8.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.8.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.8.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.8.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.8.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.8.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.8.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.8.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.8.norm2.weight, torch.Size([768])
```

retriever.query_encoder.layers.8.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.9.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.9.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.9.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.9.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.9.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.9.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.9.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.9.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.9.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.9.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.9.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.9.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.9.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.9.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.9.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.9.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.10.attention.q_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.10.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.10.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.10.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.10.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.10.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.10.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.10.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.10.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.10.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.10.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.10.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.10.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.10.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.10.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.10.norm2.bias, torch.Size([768])
retriever.query_encoder.layers.11.attention.q_lin.weight,

```
torch.Size([768, 768])
retriever.query_encoder.layers.11.attention.q_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.11.attention.k_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.11.attention.k_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.11.attention.v_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.11.attention.v_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.11.attention.out_lin.weight,
torch.Size([768, 768])
retriever.query_encoder.layers.11.attention.out_lin.bias,
torch.Size([768])
retriever.query_encoder.layers.11.norm1.weight, torch.Size([768])
retriever.query_encoder.layers.11.norm1.bias, torch.Size([768])
retriever.query_encoder.layers.11.ffn.lin1.weight, torch.Size([3072,
768])
retriever.query_encoder.layers.11.ffn.lin1.bias, torch.Size([3072])
retriever.query_encoder.layers.11.ffn.lin2.weight, torch.Size([768,
3072])
retriever.query_encoder.layers.11.ffn.lin2.bias, torch.Size([768])
retriever.query_encoder.layers.11.norm2.weight, torch.Size([768])
retriever.query_encoder.layers.11.norm2.bias, torch.Size([768])
seq2seq_decoder.embeddings.weight, torch.Size([50264, 1024])
seq2seq_decoder.norm_embeddings.weight, torch.Size([1024])
seq2seq_decoder.norm_embeddings.bias, torch.Size([1024])
seq2seq_decoder.position_embeddings.weight, torch.Size([1024, 1024])
seq2seq_decoder.layers.0.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.0.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.0.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.0.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.0.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.0.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.0.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.0.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.encoder_attention.k_lin.bias,
```

```
torch.Size([1024])
seq2seq_decoder.layers.0.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.0.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.0.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.0.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.0.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.0.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.0.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.0.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.0.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.0.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.0.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.1.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.1.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.1.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.1.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.1.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.1.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.1.ffn.lin1.weight, torch.Size([4096, 1024])
```

```
seq2seq_decoder.layers.1.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.1.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.1.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.1.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.1.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.2.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.2.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.2.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.2.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.2.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.2.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.2.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.2.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.2.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.2.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.2.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.2.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.3.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.3.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.self_attention.k_lin.bias,
```

```
torch.Size([1024])
seq2seq_decoder.layers.3.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.3.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.3.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.3.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.3.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.3.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.3.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.3.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.3.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.3.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.3.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.3.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.3.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.3.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.3.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.3.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.3.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.4.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.4.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.4.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.4.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.4.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.4.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.4.encoder_attention.q_lin.weight,
```

torch.Size([1024, 1024])
seq2seq_decoder.layers.4.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.4.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.4.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.4.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.4.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.4.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.4.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.4.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.4.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.4.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.4.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.4.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.5.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.5.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.5.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.encoder_attention.out_lin.weight,

```
torch.Size([1024, 1024])
seq2seq_decoder.layers.5.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.5.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.5.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.5.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.5.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.5.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.5.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.5.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.5.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.6.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.6.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.6.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.6.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.6.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.6.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.6.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.6.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.6.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.6.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.6.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.6.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.7.self_attention.q_lin.weight,
```

```
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.7.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.7.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.7.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.7.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.7.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.7.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.7.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.7.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.7.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.7.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.7.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.8.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.self_attention.out_lin.weight,
```

```
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.8.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.8.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.8.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.8.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.8.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.8.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.8.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.8.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.8.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.8.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.8.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.9.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.9.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.9.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.9.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.9.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.9.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.9.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.9.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.encoder_attention.k_lin.bias,
```

```
torch.Size([1024])
seq2seq_decoder.layers.9.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.9.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.9.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.9.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.9.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.9.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.9.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.9.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.9.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.9.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.9.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.10.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.10.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.10.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.10.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.10.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.10.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.10.ffn.lin1.weight, torch.Size([4096, 1024])
```

```
seq2seq_decoder.layers.10.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.10.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.10.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.10.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.10.norm3.bias, torch.Size([1024])
seq2seq_decoder.layers.11.self_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.self_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.self_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.self_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.self_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.self_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.self_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.self_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.norm1.weight, torch.Size([1024])
seq2seq_decoder.layers.11.norm1.bias, torch.Size([1024])
seq2seq_decoder.layers.11.encoder_attention.q_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.encoder_attention.q_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.encoder_attention.k_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.encoder_attention.k_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.encoder_attention.v_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.encoder_attention.v_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.encoder_attention.out_lin.weight,
torch.Size([1024, 1024])
seq2seq_decoder.layers.11.encoder_attention.out_lin.bias,
torch.Size([1024])
seq2seq_decoder.layers.11.norm2.weight, torch.Size([1024])
seq2seq_decoder.layers.11.norm2.bias, torch.Size([1024])
seq2seq_decoder.layers.11.ffn.lin1.weight, torch.Size([4096, 1024])
seq2seq_decoder.layers.11.ffn.lin1.bias, torch.Size([4096])
seq2seq_decoder.layers.11.ffn.lin2.weight, torch.Size([1024, 4096])
seq2seq_decoder.layers.11.ffn.lin2.bias, torch.Size([1024])
seq2seq_decoder.layers.11.norm3.weight, torch.Size([1024])
seq2seq_decoder.layers.11.norm3.bias, torch.Size([1024])
```