Training 8k data points

```
Prepare tokenizer, pretrained model and optimizer - add special tokens
for fine-tuning

Model: <class
'transformers.models.openai.modeling_openai.OpenAIGPTLMHeadModel'>
Config: <class
'transformers.models.openai.configuration_openai.OpenAIGPTConfig'>
Tokenizer: <class
'transformers.models.bert.tokenization_bert.BertTokenizer'>

Using pretrained model

Downloading: 100% 232k/232k [00:00<00:00, 3.03MB/s]
Downloading: 100% 28.0/28.0 [00:00<00:00, 24.4kB/s]
Downloading: 100% 570/570 [00:00<00:00, 457kB/s]
You are using a model of type bert to instantiate a model of type
openai-gpt. This is not supported for all configurations of models and
can yield errors.
Downloading: 100% 440M/440M [00:06<00:00, 66.4MB/s]
Some weights of the model checkpoint at bert-base-uncased were not used
when initializing OpenAIGPTLMHeadModel:
['bert.encoder.layer.11.attention.self.key.bias',
'bert.encoder.layer.5.attention.self.key.weight',
'bert.encoder.layer.1.attention.self.key.bias',
'bert.encoder.layer.5.attention.output.dense.bias',
'bert.encoder.layer.0.output.dense.bias',
'bert.encoder.layer.1.attention.self.value.weight',
'bert.encoder.layer.9.output.LayerNorm.weight',
'bert.encoder.layer.11.attention.self.value.weight',
'bert.encoder.layer.0.intermediate.dense.weight',
'bert.encoder.layer.2.attention.self.value.bias',
'bert.encoder.layer.2.intermediate.dense.bias',
'bert.encoder.layer.5.attention.output.dense.weight',
'bert.encoder.layer.5.attention.self.value.bias',
'bert.encoder.layer.11.output.LayerNorm.weight',
'bert.encoder.layer.6.attention.self.value.bias',
'bert.encoder.layer.1.attention.output.dense.weight',
'bert.encoder.layer.2.attention.self.query.bias',
'bert.encoder.layer.4.attention.self.query.bias',
'cls.predictions.transform.LayerNorm.bias',
'bert.encoder.layer.11.intermediate.dense.bias',
'bert.encoder.layer.4.attention.output.LayerNorm.weight',
'bert.encoder.layer.8.attention.output.dense.bias',
'bert.encoder.layer.10.attention.output.LayerNorm.weight',
'bert.encoder.layer.0.attention.output.dense.weight',
'bert.encoder.layer.6.output.LayerNorm.bias',
'bert.encoder.layer.11.attention.self.query.bias',
```

```
'bert.encoder.layer.11.attention.self.key.weight',
'bert.encoder.layer.9.attention.self.query.weight',
'bert.encoder.layer.7.intermediate.dense.weight',
'bert.embeddings.position_embeddings.weight',
'bert.encoder.layer.3.output.dense.bias',
'bert.encoder.layer.7.attention.self.query.bias',
'bert.encoder.layer.6.intermediate.dense.weight',
'bert.encoder.layer.1.attention.self.key.weight',
'bert.encoder.layer.3.attention.self.value.weight',
'bert.encoder.layer.4.attention.output.LayerNorm.bias',
'bert.encoder.layer.10.attention.self.query.weight',
'bert.encoder.layer.10.attention.self.key.bias',
'bert.embeddings.token_type_embeddings.weight',
'bert.encoder.layer.10.attention.output.dense.weight',
'bert.encoder.layer.3.attention.output.dense.bias',
'bert.encoder.layer.6.attention.self.key.bias',
'bert.encoder.layer.7.attention.output.dense.bias',
'bert.encoder.layer.9.output.dense.bias',
'bert.encoder.layer.1.attention.output.dense.bias',
'bert.encoder.layer.0.attention.self.key.weight',
'bert.encoder.layer.2.attention.self.query.weight',
'bert.encoder.layer.8.output.dense.bias',
'bert.encoder.layer.1.attention.output.LayerNorm.weight',
'bert.encoder.layer.9.attention.self.key.weight',
'bert.encoder.layer.2.output.dense.weight',
'bert.embeddings.LayerNorm.bias',
'bert.encoder.layer.6.attention.self.value.weight',
'bert.encoder.layer.4.attention.self.value.weight',
'bert.encoder.layer.11.attention.output.dense.bias',
'bert.encoder.layer.10.output.dense.bias',
'bert.encoder.layer.0.attention.self.value.weight',
'bert.encoder.layer.1.attention.self.query.bias',
'bert.encoder.layer.10.intermediate.dense.bias',
'bert.encoder.layer.8.intermediate.dense.weight',
'bert.encoder.layer.5.output.dense.weight',
'bert.encoder.layer.8.attention.self.value.weight',
'bert.encoder.layer.9.attention.output.LayerNorm.weight',
'bert.encoder.layer.7.attention.output.LayerNorm.weight',
'bert.encoder.layer.5.attention.self.query.bias',
'bert.encoder.layer.4.intermediate.dense.bias',
'bert.encoder.layer.1.output.LayerNorm.weight',
'bert.encoder.layer.5.attention.self.value.weight',
'bert.encoder.layer.7.intermediate.dense.bias',
'bert.encoder.layer.3.output.dense.weight',
'bert.encoder.layer.11.output.dense.weight',
'bert.encoder.layer.2.attention.self.value.weight',
'bert.encoder.layer.6.attention.self.key.weight',
'bert.encoder.layer.1.intermediate.dense.weight',
'bert.encoder.layer.3.attention.self.value.bias',
'bert.encoder.layer.5.attention.output.LayerNorm.weight',
'bert.encoder.layer.3.attention.self.key.bias',
```

```
'bert.encoder.layer.4.attention.self.query.weight',
'bert.encoder.layer.9.attention.self.query.bias',
'bert.encoder.layer.0.attention.self.key.bias',
'bert.encoder.layer.0.attention.output.LayerNorm.bias',
'bert.encoder.layer.11.intermediate.dense.weight',
'bert.encoder.layer.4.attention.self.key.weight',
'bert.encoder.layer.7.attention.self.value.bias',
'bert.encoder.layer.3.output.LayerNorm.weight',
'bert.encoder.layer.4.output.dense.bias',
'bert.encoder.layer.0.attention.self.query.bias',
'bert.encoder.layer.2.intermediate.dense.weight',
'bert.encoder.layer.8.attention.output.dense.weight',
'bert.encoder.layer.9.attention.self.key.bias',
'bert.encoder.layer.9.attention.output.dense.bias',
'bert.encoder.layer.0.intermediate.dense.bias',
'bert.encoder.layer.10.output.LayerNorm.bias',
'bert.encoder.layer.8.attention.output.LayerNorm.weight',
'bert.encoder.layer.9.attention.output.LayerNorm.bias',
'bert.encoder.layer.1.output.dense.bias',
'bert.encoder.layer.5.attention.output.LayerNorm.bias',
'bert.encoder.layer.2.attention.self.key.bias',
'bert.encoder.layer.9.intermediate.dense.bias',
'bert.encoder.layer.6.attention.self.query.weight',
'bert.encoder.layer.9.output.LayerNorm.bias',
'bert.encoder.layer.3.attention.output.dense.weight',
'bert.encoder.layer.9.intermediate.dense.weight',
'bert.encoder.layer.4.output.LayerNorm.bias',
'bert.encoder.layer.3.attention.self.query.bias',
'cls.predictions.bias',
'bert.encoder.layer.10.attention.self.value.bias',
'bert.encoder.layer.8.intermediate.dense.bias',
'bert.encoder.layer.4.attention.output.dense.weight',
'bert.encoder.layer.8.attention.self.query.weight',
'bert.encoder.layer.6.attention.output.dense.weight',
'bert.embeddings.LayerNorm.weight',
'bert.encoder.layer.0.attention.output.dense.bias',
'cls.predictions.transform.LayerNorm.weight',
'cls.seq_relationship.weight',
'bert.encoder.layer.5.intermediate.dense.weight',
'cls.seq_relationship.bias',
'bert.encoder.layer.7.output.LayerNorm.bias',
'bert.encoder.layer.0.output.dense.weight',
'bert.encoder.layer.4.attention.self.value.bias',
'bert.encoder.layer.10.attention.self.key.weight',
'bert.encoder.layer.5.intermediate.dense.bias',
'bert.encoder.layer.6.attention.output.LayerNorm.bias',
'bert.encoder.layer.7.attention.output.LayerNorm.bias',
'bert.encoder.layer.8.output.LayerNorm.bias',
'bert.embeddings.word_embeddings.weight',
'bert.encoder.layer.3.output.LayerNorm.bias',
'bert.encoder.layer.2.attention.output.LayerNorm.bias',
```

```
'bert.encoder.layer.9.attention.output.dense.weight',
'bert.encoder.layer.10.output.dense.weight',
'cls.predictions.decoder.weight',
'bert.encoder.layer.2.output.dense.bias',
'bert.encoder.layer.5.attention.self.query.weight',
'bert.encoder.layer.2.attention.output.dense.weight',
'bert.encoder.layer.6.output.dense.weight', 'bert.pooler.dense.bias',
'bert.encoder.layer.3.attention.output.LayerNorm.weight',
'bert.encoder.layer.11.attention.self.value.bias',
'bert.pooler.dense.weight',
'bert.encoder.layer.2.output.LayerNorm.bias',
'bert.encoder.layer.10.attention.output.dense.bias',
'bert.encoder.layer.6.output.dense.bias',
'bert.encoder.layer.2.attention.output.LayerNorm.weight',
'bert.encoder.layer.7.output.dense.bias',
'bert.encoder.layer.9.attention.self.value.weight',
'bert.encoder.layer.5.output.dense.bias',
'bert.encoder.layer.7.attention.self.value.weight',
'bert.encoder.layer.10.attention.self.value.weight',
'bert.encoder.layer.8.attention.output.LayerNorm.bias',
'bert.encoder.layer.7.output.dense.weight',
'bert.encoder.layer.10.attention.output.LayerNorm.bias',
'bert.encoder.layer.8.attention.self.query.bias',
'bert.encoder.layer.11.output.LayerNorm.bias',
'bert.encoder.layer.3.intermediate.dense.bias',
'bert.encoder.layer.3.attention.output.LayerNorm.bias',
'bert.encoder.layer.8.output.LayerNorm.weight',
'bert.encoder.layer.1.output.dense.weight',
'bert.encoder.layer.4.output.dense.weight',
'bert.encoder.layer.8.output.dense.weight',
'bert.encoder.layer.6.attention.output.dense.bias',
'bert.encoder.layer.1.attention.output.LayerNorm.bias',
'bert.encoder.layer.10.intermediate.dense.weight',
'bert.encoder.layer.1.attention.self.value.bias',
'bert.encoder.layer.11.attention.output.LayerNorm.bias',
'bert.encoder.layer.11.output.dense.bias',
'bert.encoder.layer.4.attention.output.dense.bias',
'bert.encoder.layer.10.attention.self.query.bias',
'bert.encoder.layer.7.output.LayerNorm.weight',
'bert.encoder.layer.2.attention.self.key.weight',
'bert.encoder.layer.4.output.LayerNorm.weight',
'bert.encoder.layer.11.attention.output.dense.weight',
'bert.encoder.layer.3.attention.self.key.weight',
'bert.encoder.layer.7.attention.output.dense.weight',
'bert.encoder.layer.8.attention.self.key.weight',
'bert.encoder.layer.6.intermediate.dense.bias',
'bert.encoder.layer.6.output.LayerNorm.weight',
'bert.encoder.layer.10.output.LayerNorm.weight',
'bert.encoder.layer.4.attention.self.key.bias',
'bert.encoder.layer.1.attention.self.query.weight',
'bert.encoder.layer.7.attention.self.query.weight',
```

```
'bert.encoder.layer.1.output.LayerNorm.bias',
'bert.encoder.layer.7.attention.self.key.weight',
'bert.encoder.layer.5.output.LayerNorm.bias',
'bert.encoder.layer.0.attention.self.value.bias',
'bert.encoder.layer.3.intermediate.dense.weight',
'bert.encoder.layer.0.output.LayerNorm.bias',
'bert.encoder.layer.5.output.LayerNorm.weight',
'bert.encoder.layer.9.attention.self.value.bias',
'bert.encoder.layer.9.output.dense.weight',
'bert.encoder.layer.5.attention.self.key.bias',
'bert.encoder.layer.8.attention.self.key.bias',
'bert.encoder.layer.0.output.LayerNorm.weight',
'bert.encoder.layer.0.attention.self.query.weight',
'bert.encoder.layer.8.attention.self.value.bias',
'bert.encoder.layer.1.intermediate.dense.bias',
'bert.encoder.layer.6.attention.output.LayerNorm.weight',
'bert.encoder.layer.6.attention.self.query.bias',
'bert.encoder.layer.3.attention.self.query.weight',
'bert.encoder.layer.11.attention.output.LayerNorm.weight',
'bert.encoder.layer.2.attention.output.dense.bias',
'bert.encoder.layer.7.attention.self.key.bias',
'bert.encoder.layer.4.intermediate.dense.weight',
'bert.encoder.layer.11.attention.self.query.weight',
'cls.predictions.transform.dense.bias',
'bert.encoder.layer.2.output.LayerNorm.weight',
'cls.predictions.transform.dense.weight',
'bert.encoder.layer.0.attention.output.LayerNorm.weight']
- This IS expected if you are initializing OpenAIGPTLMHeadModel from
the checkpoint of a model trained on another task or with another
architecture (e.g. initializing a BertForSequenceClassification model
from a BertForPreTraining model).
- This IS NOT expected if you are initializing OpenAIGPTLMHeadModel
from the checkpoint of a model that you expect to be exactly identical
(initializing a BertForSequenceClassification model from a
BertForSequenceClassification model).
Some weights of OpenAIGPTLMHeadModel were not initialized from the
model checkpoint at bert-base-uncased and are newly initialized:
['h.1.mlp.c_fc.bias', 'h.5.ln_2.bias', 'h.3.mlp.c_proj.weight',
'h.7.attn.c_attn.bias', 'h.10.mlp.c_fc.weight', 'h.6.mlp.c_fc.weight',
'h.1.ln_1.weight', 'h.10.attn.c_attn.weight', 'h.9.attn.c_attn.weight',
'h.6.attn.bias', 'h.11.attn.c_attn.weight', 'h.1.ln_2.weight',
'h.5.attn.bias', 'h.7.mlp.c_fc.bias', 'h.1.attn.c_proj.weight',
'h.2.attn.bias', 'h.4.ln_2.bias', 'h.0.attn.c_proj.weight',
'h.9.ln_2.bias', 'h.10.attn.bias', 'h.4.attn.c_proj.weight',
'h.2.ln_1.bias', 'h.9.ln_1.bias', 'h.0.attn.c_attn.bias',
'h.10.ln_2.bias', 'h.7.ln_1.bias', 'h.0.attn.c_attn.weight',
'h.2.attn.c_proj.bias', 'h.6.mlp.c_proj.bias', 'h.5.ln_2.weight',
'h.10.attn.c_proj.bias', 'h.4.mlp.c_fc.bias', 'h.8.mlp.c_proj.bias',
'h.10.mlp.c_proj.bias', 'h.9.mlp.c_proj.weight',
'h.7.mlp.c_proj.weight', 'h.4.attn.c_attn.weight',
'h.5.mlp.c_proj.weight', 'h.8.mlp.c_fc.weight', 'h.3.mlp.c_fc.weight',
```

```
'h.11.attn.bias', 'h.1.ln_2.bias', 'h.6.mlp.c_proj.weight',
'h.3.ln_1.weight', 'h.3.ln_2.bias', 'h.6.attn.c_attn.bias',
'h.5.attn.c_attn.weight', 'h.8.attn.bias', 'h.3.attn.bias',
'h.6.mlp.c_fc.bias', 'h.10.attn.c_proj.weight', 'h.1.mlp.c_fc.weight',
'h.4.attn.c_attn.bias', 'h.9.ln_2.weight', 'h.0.mlp.c_fc.bias',
'h.0.attn.c_proj.bias', 'h.5.mlp.c_fc.weight', 'h.11.mlp.c_proj.bias',
'h.4.ln_1.weight', 'h.0.ln_2.bias', 'h.4.mlp.c_proj.bias',
'h.11.mlp.c_proj.weight', 'h.5.mlp.c_proj.bias', 'h.4.ln_2.weight',
'h.5.mlp.c_fc.bias', 'h.11.ln_1.weight', 'h.9.attn.c_proj.bias',
'h.5.attn.c_proj.weight', 'h.1.attn.c_attn.bias', 'h.6.ln_2.bias',
'h.11.ln_2.weight', 'h.6.ln_2.weight', 'h.10.ln_1.weight',
'h.5.ln_1.weight', 'h.8.attn.c_attn.weight', 'h.3.ln_2.weight',
'h.8.attn.c_proj.bias', 'h.0.attn.bias', 'h.3.ln_1.bias',
'h.11.attn.c_proj.weight', 'h.11.ln_1.bias', 'h.2.mlp.c_proj.bias',
'h.8.ln_1.bias', 'h.1.mlp.c_proj.bias', 'h.11.mlp.c_fc.bias',
'h.0.mlp.c_proj.bias', 'h.1.attn.bias', 'h.5.attn.c_attn.bias',
'h.9.attn.c_proj.weight', 'h.5.attn.c_proj.bias', 'h.2.mlp.c_fc.bias',
'h.8.attn.c_attn.bias', 'h.10.mlp.c_fc.bias', 'h.0.ln_1.bias',
'h.4.mlp.c_fc.weight', 'h.10.ln_2.weight', 'h.4.attn.bias',
'h.7.mlp.c_proj.bias', 'h.9.mlp.c_proj.bias', 'h.8.ln_2.bias',
'h.0.ln_1.weight', 'h.3.attn.c_attn.weight', 'h.7.attn.c_proj.bias',
'h.3.attn.c_attn.bias', 'h.3.mlp.c_fc.bias', 'h.3.mlp.c_proj.bias',
'h.7.ln_2.weight', 'h.8.ln_2.weight', 'h.11.mlp.c_fc.weight',
'h.1.ln_1.bias', 'h.5.ln_1.bias', 'h.9.attn.c_attn.bias',
'tokens_embed.weight', 'h.3.attn.c_proj.bias', 'h.7.attn.bias',
'h.6.attn.c_attn.weight', 'h.7.mlp.c_fc.weight',
'h.2.mlp.c_proj.weight', 'h.2.ln_2.bias', 'h.11.ln_2.bias',
'h.11.attn.c_proj.bias', 'h.1.attn.c_attn.weight', 'h.2.ln_2.weight',
'h.3.attn.c_proj.weight', 'h.8.ln_1.weight', 'h.6.ln_1.weight',
'h.8.mlp.c_proj.weight', 'h.11.attn.c_attn.bias',
'h.0.mlp.c_fc.weight', 'lm_head.weight', 'h.2.attn.c_proj.weight',
'h.2.attn.c_attn.weight', 'h.0.mlp.c_proj.weight',
'h.1.mlp.c_proj.weight', 'h.9.attn.bias', 'h.0.ln_2.weight',
'h.6.attn.c_proj.bias', 'h.7.attn.c_proj.weight', 'h.10.ln_1.bias',
'h.4.mlp.c_proj.weight', 'h.9.mlp.c_fc.weight',
'h.10.mlp.c_proj.weight', 'h.7.attn.c_attn.weight', 'h.7.ln_1.weight',
'h.10.attn.c_attn.bias', 'h.7.ln_2.bias', 'h.1.attn.c_proj.bias',
'h.9.mlp.c_fc.bias', 'h.2.attn.c_attn.bias', 'h.2.mlp.c_fc.weight',
'h.8.attn.c_proj.weight', 'h.4.attn.c_proj.bias', 'h.4.ln_1.bias',
'h.2.ln_1.weight', 'h.9.ln_1.weight', 'positions_embed.weight',
'h.6.ln_1.bias', 'h.6.attn.c_proj.weight', 'h.8.mlp.c_fc.bias']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

Tokenizer Length: 30527

/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:310
: FutureWarning: This implementation of AdamW is deprecated and will be
removed in a future version. Use the PyTorch implementation
torch.optim.AdamW instead, or set `no_deprecation_warning=True` to
disable this warning
```

```
  FutureWarning,

Optimizer: AdamW (
Parameter Group 0
    betas: (0.9, 0.999)
    correct_bias: True
    eps: 1e-06
    initial_lr: 0.001
    lr: 0.001
    weight_decay: 0.0
)

Build train and validation dataloaders

/usr/local/lib/python3.7/dist-packages/ignite/handlers/checkpoint.py:99
3: UserWarning: Argument save_interval is deprecated and should be
None. This argument will be removed in 0.5.0.Please, use events
filtering instead, e.g. Events.ITERATION_STARTED(every=1000)
  warnings.warn(msg)

Validation: {'average_nll': 3.2287888292466014,
 'average_ppl': 25.249057524101882,
 'nll': 3.2287888292466014}
Epoch [1/3]: 100% 500/500 [1:47:22<00:00, 12.91s/it, loss=0.0556,
lr=0.0007]

Validation: {'average_nll': 2.3481067102867064,
 'average_ppl': 10.465736284578322,
 'nll': 2.3481067102867064}
Epoch [2/3]: 100% 500/500 [1:47:14<00:00, 12.89s/it, loss=0.0386,
lr=0.000367]

Validation: {'average_nll': 2.190880850855461,
 'average_ppl': 8.943087172240597,
 'nll': 2.190880850855461}
Epoch [3/3]: 100% 500/500 [1:50:05<00:00, 13.24s/it, loss=0.0346,
lr=3.33e-5]
```

---

lr = 0.000000001, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:21<00:00, 13.65s/it, loss=0.133, lr=9.81e-10]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
```

```
Epoch [2/3]: 100% 7/7 [01:22<00:00, 13.79s/it, loss=0.135, lr=6.47e-10]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:31<00:00, 15.29s/it, loss=0.134, lr=3.14e-10]
```

lr = 0.000001, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:19<00:00, 13.20s/it, loss=0.133, lr=9.81e-7]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [2/3]: 100% 7/7 [01:23<00:00, 13.92s/it, loss=0.135, lr=6.47e-7]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:18<00:00, 13.11s/it, loss=0.134, lr=3.14e-7]
```

lr = 0.0001, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:21<00:00, 13.54s/it, loss=0.133, lr=9.81e-5]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [2/3]: 100% 7/7 [01:26<00:00, 14.48s/it, loss=0.135, lr=6.47e-5]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:22<00:00, 13.76s/it, loss=0.134, lr=3.14e-5]
```

lr = 0.001, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:23<00:00, 13.96s/it, loss=0.133, lr=0.000981]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
```

```
Epoch [2/3]: 100% 7/7 [01:18<00:00, 13.08s/it, loss=0.135, lr=0.000647]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:23<00:00, 13.93s/it, loss=0.134, lr=0.000314]
```

lr = 0.01, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:27<00:00, 14.56s/it, loss=0.133, lr=0.00981]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [2/3]: 100% 7/7 [01:24<00:00, 14.04s/it, loss=0.135, lr=0.00647]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:22<00:00, 13.81s/it, loss=0.134, lr=0.00314]
```

lr = 0.1, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:19<00:00, 13.28s/it, loss=0.133, lr=0.0981]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [2/3]: 100% 7/7 [01:28<00:00, 14.72s/it, loss=0.135, lr=0.0647]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:27<00:00, 14.66s/it, loss=0.134, lr=0.0314]
```

lr = 0.5, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:20<00:00, 13.44s/it, loss=0.133, lr=0.49]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
```

```
Epoch [2/3]: 100% 7/7 [01:20<00:00, 13.48s/it, loss=0.135, lr=0.324]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:25<00:00, 14.27s/it, loss=0.134, lr=0.157]
```

lr = 0.9, 100 datapoints

```
Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [1/3]: 100% 7/7 [01:23<00:00, 13.87s/it, loss=0.133, lr=0.883]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [2/3]: 100% 7/7 [01:19<00:00, 13.26s/it, loss=0.135, lr=0.583]

Validation: {'average_nll': 8.39444200351494,
 'average_ppl': 4422.418527160187,
 'nll': 8.39444200351494}
Epoch [3/3]: 100% 7/7 [01:24<00:00, 14.12s/it, loss=0.134, lr=0.283]
```

```
!python train_moban.py --pretrained --model_checkpoint
"bert-base-uncased" --train_path ../data/raw/train2.json --valid_path
../data/raw/train2.json --scheduler linear --num_workers 2
--valid_steps 3 --lr 0.001 --gpt2
```

```
Model: <class 'transformers.models.gpt2.modeling_gpt2.GPT2LMHeadModel'>
Config: <class
'transformers.models.gpt2.configuration_gpt2.GPT2Config'>
Tokenizer: <class
'transformers.models.bert.tokenization_bert.BertTokenizer'>

Validation: {'average_nll': 10.651066838752197,
 'average_ppl': 42237.631262935116,
 'nll': 10.651066838752197}
Epoch [1/3]: 100% 7/7 [01:11<00:00, 11.92s/it, loss=0.165, lr=0.000981]

Validation: {'average_nll': 10.651066838752197,
 'average_ppl': 42237.631262935116,
 'nll': 10.651066838752197}
Epoch [2/3]: 100% 7/7 [01:10<00:00, 11.68s/it, loss=0.166, lr=0.000647]

Validation: {'average_nll': 10.651066838752197,
 'average_ppl': 42237.631262935116,
 'nll': 10.651066838752197}
```

```
Epoch [3/3]: 100% 7/7 [01:10<00:00, 11.83s/it, loss=0.165, lr=0.000314]
```

```
!python train_moban.py --pretrained --model_checkpoint "gpt2"
--train_path ../data/raw/train2.json --valid_path
../data/raw/train2.json --scheduler linear --num_workers 2
--valid_steps 3 --lr 0.001 --gpt2
```

```
The cache for model files in Transformers v4.22.0 has been updated.
Migrating your old cache. This is a one-time only operation. You can
interrupt this and resume the migration later on by calling
`transformers.utils.move_cache()`.
Moving 0 files to the new cache system
0it [00:00, ?it/s]
Prepare tokenizer, pretrained model and optimizer - add special tokens
for fine-tuning

Model: <class 'transformers.models.gpt2.modeling_gpt2.GPT2LMHeadModel'>
Config: <class
'transformers.models.gpt2.configuration_gpt2.GPT2Config'>
Tokenizer: <class
'transformers.models.gpt2.tokenization_gpt2.GPT2Tokenizer'>

Using pretrained model


Tokenizer Length: 50262

/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:310
: FutureWarning: This implementation of AdamW is deprecated and will be
removed in a future version. Use the PyTorch implementation
torch.optim.AdamW instead, or set `no_deprecation_warning=True` to
disable this warning
  FutureWarning,

Optimizer: AdamW (
Parameter Group 0
    betas: (0.9, 0.999)
    correct_bias: True
    eps: 1e-06
    initial_lr: 0.001
    lr: 0.001
    weight_decay: 0.0
)

Build train and validation dataloaders

/usr/local/lib/python3.7/dist-packages/ignite/handlers/checkpoint.py:99
3: UserWarning: Argument save_interval is deprecated and should be
```

```
None. This argument will be removed in 0.5.0.Please, use events
filtering instead, e.g. Events.ITERATION_STARTED(every=1000)
  warnings.warn(msg)

Validation: {'average_nll': 82.87623708073323,
 'average_ppl': 9.833145039231839e+35,
 'nll': 82.87623708073323}
Epoch [1/3]: 100% 7/7 [13:16<00:00, 132.67s/it, loss=1.18, lr=0.000981]

Validation: {'average_nll': 82.87623708073323,
 'average_ppl': 9.833145039231839e+35,
 'nll': 82.87623708073323}
Epoch [2/3]: 100% 7/7 [13:36<00:00, 136.12s/it, loss=1.17, lr=0.000647]

Validation: {'average_nll': 82.87623708073323,
 'average_ppl': 9.833145039231839e+35,
 'nll': 82.87623708073323}
Epoch [3/3]: 100% 7/7 [13:43<00:00, 137.18s/it, loss=1.15, lr=0.000314]

--------------------------------------------------------------------------
```

Training on English dataset : *bert-base-uncased*

```
Prepare tokenizer, pretrained model and optimizer - add special tokens
for fine-tuning

Model: <class
'transformers.models.openai.modeling_openai.OpenAIGPTLMHeadModel'>
Config: <class
'transformers.models.openai.configuration_openai.OpenAIGPTConfig'>
Tokenizer: <class
'transformers.models.bert.tokenization_bert.BertTokenizer'>

Using pretrained model

Downloading: 100% 232k/232k [00:00<00:00, 1.24MB/s]
Downloading: 100% 28.0/28.0 [00:00<00:00, 16.1kB/s]
Downloading: 100% 570/570 [00:00<00:00, 447kB/s]
You are using a model of type bert to instantiate a model of type
openai-gpt. This is not supported for all configurations of models and
can yield errors.
Downloading: 100% 440M/440M [00:07<00:00, 57.0MB/s]
Some weights of the model checkpoint at bert-base-uncased were not used
when initializing OpenAIGPTLMHeadModel:
['bert.embeddings.word_embeddings.weight',
'bert.encoder.layer.5.attention.self.key.weight',
'bert.embeddings.LayerNorm.bias',
'bert.encoder.layer.11.intermediate.dense.weight',
'bert.encoder.layer.10.attention.output.dense.weight',
'cls.predictions.transform.LayerNorm.bias',
'cls.seq_relationship.weight',
```

'bert.encoder.layer.11.attention.self.key.bias',
'bert.encoder.layer.1.attention.self.key.weight',
'bert.encoder.layer.6.attention.output.dense.weight',
'bert.encoder.layer.7.output.dense.weight',
'bert.encoder.layer.4.attention.self.key.weight',
'bert.encoder.layer.5.attention.output.dense.weight',
'bert.encoder.layer.5.output.LayerNorm.bias',
'bert.encoder.layer.5.attention.output.LayerNorm.bias',
'bert.encoder.layer.2.attention.self.query.weight',
'bert.encoder.layer.2.attention.output.LayerNorm.bias',
'bert.encoder.layer.0.intermediate.dense.weight',
'bert.encoder.layer.1.attention.output.dense.bias',
'bert.encoder.layer.7.attention.self.key.weight',
'bert.encoder.layer.11.attention.output.LayerNorm.bias',
'bert.encoder.layer.10.attention.self.query.bias',
'bert.encoder.layer.1.intermediate.dense.bias',
'bert.encoder.layer.9.attention.self.key.bias',
'bert.encoder.layer.5.output.LayerNorm.weight',
'bert.encoder.layer.6.intermediate.dense.bias',
'bert.encoder.layer.6.attention.self.key.weight',
'bert.embeddings.LayerNorm.weight',
'bert.encoder.layer.3.attention.self.key.bias',
'bert.encoder.layer.8.output.dense.bias',
'bert.encoder.layer.3.output.LayerNorm.weight',
'bert.encoder.layer.2.attention.self.query.bias',
'bert.encoder.layer.5.attention.self.value.bias',
'bert.encoder.layer.6.intermediate.dense.weight',
'bert.encoder.layer.8.attention.self.value.weight',
'bert.encoder.layer.11.attention.self.query.weight',
'bert.encoder.layer.4.attention.output.dense.weight',
'bert.encoder.layer.4.attention.self.query.weight',
'bert.encoder.layer.7.attention.output.dense.bias',
'bert.encoder.layer.9.attention.self.query.weight',
'bert.encoder.layer.9.output.dense.bias',
'bert.encoder.layer.4.attention.output.LayerNorm.bias',
'bert.pooler.dense.bias', 'bert.encoder.layer.10.output.dense.bias',
'bert.encoder.layer.3.attention.self.value.weight',
'bert.encoder.layer.4.output.dense.bias',
'bert.encoder.layer.6.attention.self.value.bias',
'bert.encoder.layer.11.attention.output.dense.weight',
'bert.encoder.layer.4.attention.self.value.weight',
'bert.encoder.layer.8.intermediate.dense.bias',
'bert.encoder.layer.2.attention.self.key.bias',
'bert.encoder.layer.2.output.dense.bias',
'bert.encoder.layer.0.output.LayerNorm.bias',
'bert.encoder.layer.9.output.LayerNorm.bias',
'bert.encoder.layer.3.output.LayerNorm.bias',
'bert.encoder.layer.7.output.dense.bias',
'bert.encoder.layer.10.attention.self.key.bias',
'bert.encoder.layer.0.attention.output.dense.bias',
'bert.encoder.layer.2.intermediate.dense.bias',

```
'bert.encoder.layer.4.output.dense.weight',
'bert.encoder.layer.5.output.dense.weight',
'bert.encoder.layer.11.attention.self.value.weight',
'bert.encoder.layer.8.attention.output.dense.weight',
'bert.encoder.layer.11.attention.output.dense.bias',
'bert.encoder.layer.3.output.dense.weight',
'bert.encoder.layer.10.output.LayerNorm.weight',
'bert.encoder.layer.11.output.dense.bias',
'bert.encoder.layer.5.attention.self.query.weight',
'bert.encoder.layer.0.attention.output.LayerNorm.bias',
'bert.encoder.layer.2.attention.output.LayerNorm.weight',
'bert.encoder.layer.8.attention.self.query.weight',
'bert.encoder.layer.8.attention.self.query.bias',
'bert.encoder.layer.8.attention.self.key.weight',
'bert.encoder.layer.7.output.LayerNorm.bias',
'bert.encoder.layer.7.attention.self.query.bias',
'cls.predictions.transform.dense.weight',
'bert.encoder.layer.0.attention.output.LayerNorm.weight',
'cls.predictions.decoder.weight',
'bert.encoder.layer.8.attention.output.LayerNorm.weight',
'bert.encoder.layer.9.attention.self.key.weight',
'bert.encoder.layer.6.attention.self.query.weight',
'bert.encoder.layer.2.attention.output.dense.weight',
'bert.encoder.layer.1.output.LayerNorm.bias',
'bert.encoder.layer.1.attention.self.value.bias',
'bert.encoder.layer.0.output.dense.weight',
'bert.encoder.layer.1.attention.self.query.bias',
'bert.encoder.layer.1.attention.output.LayerNorm.weight',
'bert.encoder.layer.11.output.LayerNorm.bias',
'bert.encoder.layer.6.output.LayerNorm.bias',
'bert.encoder.layer.9.attention.self.query.bias',
'bert.encoder.layer.2.attention.output.dense.bias',
'bert.encoder.layer.6.attention.self.value.weight',
'bert.encoder.layer.6.output.dense.bias',
'bert.encoder.layer.3.intermediate.dense.bias',
'bert.encoder.layer.7.attention.self.value.weight',
'bert.encoder.layer.1.attention.output.LayerNorm.bias',
'bert.pooler.dense.weight',
'bert.encoder.layer.8.output.LayerNorm.bias',
'bert.encoder.layer.5.attention.self.value.weight',
'bert.encoder.layer.8.output.LayerNorm.weight',
'bert.encoder.layer.3.attention.self.key.weight',
'bert.encoder.layer.4.attention.self.query.bias',
'bert.encoder.layer.7.attention.output.LayerNorm.weight',
'bert.encoder.layer.0.attention.self.value.bias',
'bert.encoder.layer.2.attention.self.value.bias',
'bert.encoder.layer.11.output.LayerNorm.weight',
'bert.encoder.layer.10.attention.self.value.weight',
'bert.encoder.layer.10.attention.self.value.bias',
'bert.encoder.layer.0.attention.self.key.bias',
'bert.encoder.layer.9.intermediate.dense.bias',
```

```
'bert.encoder.layer.1.attention.self.value.weight',
'bert.encoder.layer.1.output.dense.weight',
'bert.encoder.layer.0.attention.self.query.weight',
'bert.encoder.layer.9.attention.self.value.bias',
'bert.encoder.layer.10.attention.self.key.weight',
'bert.encoder.layer.0.output.dense.bias',
'bert.encoder.layer.5.attention.output.LayerNorm.weight',
'bert.encoder.layer.5.output.dense.bias',
'bert.encoder.layer.10.intermediate.dense.bias',
'bert.encoder.layer.4.attention.output.dense.bias',
'bert.encoder.layer.3.attention.output.dense.bias',
'bert.encoder.layer.0.attention.output.dense.weight',
'bert.encoder.layer.3.attention.output.LayerNorm.weight',
'bert.encoder.layer.8.attention.self.key.bias',
'bert.encoder.layer.6.attention.self.key.bias',
'bert.encoder.layer.8.output.dense.weight',
'bert.encoder.layer.3.attention.self.value.bias',
'bert.encoder.layer.7.attention.output.LayerNorm.bias',
'bert.encoder.layer.3.attention.self.query.bias',
'bert.encoder.layer.10.attention.output.LayerNorm.weight',
'bert.encoder.layer.4.attention.self.key.bias',
'bert.encoder.layer.4.attention.self.value.bias',
'bert.encoder.layer.5.attention.self.key.bias',
'bert.encoder.layer.7.attention.self.value.bias',
'bert.encoder.layer.3.attention.self.query.weight',
'bert.encoder.layer.0.intermediate.dense.bias',
'bert.encoder.layer.6.attention.output.dense.bias',
'bert.encoder.layer.5.intermediate.dense.weight',
'bert.encoder.layer.5.intermediate.dense.bias',
'bert.encoder.layer.11.attention.self.value.bias',
'cls.predictions.bias',
'bert.encoder.layer.3.attention.output.LayerNorm.bias',
'bert.encoder.layer.7.output.LayerNorm.weight',
'bert.encoder.layer.9.attention.output.dense.weight',
'bert.encoder.layer.9.output.dense.weight',
'cls.seq_relationship.bias',
'bert.encoder.layer.4.output.LayerNorm.weight',
'bert.encoder.layer.1.attention.output.dense.weight',
'bert.encoder.layer.3.intermediate.dense.weight',
'bert.encoder.layer.9.attention.output.LayerNorm.bias',
'bert.encoder.layer.0.output.LayerNorm.weight',
'bert.embeddings.token_type_embeddings.weight',
'bert.encoder.layer.2.output.LayerNorm.bias',
'bert.encoder.layer.10.attention.output.LayerNorm.bias',
'bert.encoder.layer.11.attention.self.key.weight',
'bert.encoder.layer.2.output.LayerNorm.weight',
'bert.encoder.layer.7.attention.output.dense.weight',
'bert.encoder.layer.2.attention.self.value.weight',
'bert.encoder.layer.6.attention.output.LayerNorm.weight',
'bert.encoder.layer.4.output.LayerNorm.bias',
'bert.encoder.layer.3.output.dense.bias',
```

'bert.encoder.layer.8.attention.self.value.bias',
'bert.encoder.layer.11.intermediate.dense.bias',
'bert.encoder.layer.5.attention.self.query.bias',
'bert.encoder.layer.1.intermediate.dense.weight',
'bert.encoder.layer.10.output.dense.weight',
'bert.encoder.layer.11.attention.output.LayerNorm.weight',
'bert.encoder.layer.0.attention.self.key.weight',
'bert.encoder.layer.9.attention.output.dense.bias',
'bert.encoder.layer.0.attention.self.query.bias',
'bert.encoder.layer.9.output.LayerNorm.weight',
'bert.encoder.layer.10.attention.self.query.weight',
'bert.encoder.layer.5.attention.output.dense.bias',
'cls.predictions.transform.LayerNorm.weight',
'bert.encoder.layer.4.intermediate.dense.bias',
'bert.encoder.layer.7.intermediate.dense.bias',
'bert.encoder.layer.9.attention.self.value.weight',
'bert.encoder.layer.4.attention.output.LayerNorm.weight',
'bert.encoder.layer.6.output.dense.weight',
'bert.encoder.layer.7.attention.self.key.bias',
'bert.encoder.layer.1.output.LayerNorm.weight',
'bert.encoder.layer.2.attention.self.key.weight',
'bert.encoder.layer.6.attention.output.LayerNorm.bias',
'bert.encoder.layer.1.attention.self.query.weight',
'bert.encoder.layer.0.attention.self.value.weight',
'bert.encoder.layer.6.output.LayerNorm.weight',
'bert.encoder.layer.9.intermediate.dense.weight',
'bert.encoder.layer.3.attention.output.dense.weight',
'bert.encoder.layer.8.attention.output.LayerNorm.bias',
'bert.encoder.layer.7.attention.self.query.weight',
'bert.embeddings.position_embeddings.weight',
'bert.encoder.layer.10.attention.output.dense.bias',
'bert.encoder.layer.10.output.LayerNorm.bias',
'bert.encoder.layer.1.attention.self.key.bias',
'bert.encoder.layer.8.attention.output.dense.bias',
'bert.encoder.layer.2.output.dense.weight',
'bert.encoder.layer.11.output.dense.weight',
'bert.encoder.layer.4.intermediate.dense.weight',
'bert.encoder.layer.2.intermediate.dense.weight',
'bert.encoder.layer.8.intermediate.dense.weight',
'bert.encoder.layer.11.attention.self.query.bias',
'bert.encoder.layer.7.intermediate.dense.weight',
'bert.encoder.layer.1.output.dense.bias',
'bert.encoder.layer.10.intermediate.dense.weight',
'bert.encoder.layer.6.attention.self.query.bias',
'cls.predictions.transform.dense.bias',
'bert.encoder.layer.9.attention.output.LayerNorm.weight']
- This IS expected if you are initializing OpenAIGPTLMHeadModel from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).

- This IS NOT expected if you are initializing OpenAIGPTLMHeadModel from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).
Some weights of OpenAIGPTLMHeadModel were not initialized from the model checkpoint at bert-base-uncased and are newly initialized:
['h.3.attn.c_proj.bias', 'h.10.ln_1.bias', 'h.1.mlp.c_proj.weight',
'h.3.mlp.c_fc.bias', 'h.1.mlp.c_proj.bias', 'h.9.mlp.c_proj.weight',
'h.11.mlp.c_fc.weight', 'h.1.mlp.c_fc.bias', 'h.7.ln_1.bias',
'h.1.attn.c_proj.bias', 'h.0.ln_2.weight', 'h.8.attn.c_proj.bias',
'h.8.ln_1.weight', 'h.11.attn.c_proj.bias', 'h.9.ln_1.weight',
'h.2.ln_1.bias', 'h.11.ln_2.bias', 'h.6.attn.c_proj.weight',
'h.0.attn.c_proj.bias', 'h.4.attn.c_proj.weight', 'h.2.ln_2.bias',
'h.9.attn.c_proj.weight', 'h.1.ln_2.bias', 'h.2.attn.c_proj.bias',
'h.7.ln_2.weight', 'h.8.mlp.c_proj.bias', 'h.1.attn.c_attn.bias',
'h.0.ln_1.bias', 'h.1.ln_2.weight', 'tokens_embed.weight',
'h.0.mlp.c_fc.weight', 'h.7.mlp.c_fc.bias', 'h.7.attn.c_proj.weight',
'h.6.attn.c_attn.bias', 'h.0.mlp.c_proj.weight', 'h.4.attn.bias',
'h.2.mlp.c_proj.weight', 'h.8.mlp.c_fc.weight',
'h.5.attn.c_proj.weight', 'h.11.attn.bias', 'h.8.ln_2.weight',
'h.6.ln_1.weight', 'h.5.attn.bias', 'h.6.mlp.c_fc.weight',
'h.2.attn.c_attn.bias', 'h.2.attn.c_proj.weight',
'h.2.mlp.c_fc.weight', 'h.8.mlp.c_fc.bias', 'h.11.ln_2.weight',
'h.3.mlp.c_fc.weight', 'h.0.attn.c_attn.bias', 'h.11.mlp.c_fc.bias',
'h.1.mlp.c_fc.weight', 'h.9.attn.c_attn.weight', 'h.3.ln_2.weight',
'h.4.attn.c_proj.bias', 'h.8.attn.c_attn.bias', 'h.0.mlp.c_proj.bias',
'h.7.ln_2.bias', 'h.10.mlp.c_proj.weight', 'h.0.mlp.c_fc.bias',
'h.4.mlp.c_fc.weight', 'h.6.attn.c_attn.weight', 'h.6.ln_2.weight',
'h.6.ln_2.bias', 'h.6.mlp.c_proj.weight', 'h.11.mlp.c_proj.bias',
'h.0.attn.bias', 'h.10.mlp.c_fc.bias', 'h.10.attn.c_attn.bias',
'h.3.ln_1.bias', 'h.3.attn.bias', 'h.4.ln_2.weight', 'h.10.attn.bias',
'h.4.mlp.c_fc.bias', 'h.9.attn.bias', 'h.6.mlp.c_proj.bias',
'h.1.ln_1.bias', 'h.3.attn.c_proj.weight', 'h.8.attn.c_proj.weight',
'h.7.attn.c_attn.bias', 'h.5.ln_1.bias', 'h.1.attn.c_proj.weight',
'h.8.ln_1.bias', 'h.2.attn.bias', 'h.3.attn.c_attn.weight',
'h.6.mlp.c_fc.bias', 'h.9.ln_1.bias', 'h.11.attn.c_proj.weight',
'h.3.ln_2.bias', 'h.2.mlp.c_proj.bias', 'h.10.attn.c_attn.weight',
'h.6.ln_1.bias', 'h.7.attn.c_attn.weight', 'h.9.mlp.c_proj.bias',
'h.0.ln_2.bias', 'h.2.mlp.c_fc.bias', 'h.4.ln_2.bias',
'positions_embed.weight', 'h.9.attn.c_attn.bias',
'h.4.attn.c_attn.weight', 'h.9.attn.c_proj.bias',
'h.5.mlp.c_proj.bias', 'lm_head.weight', 'h.1.ln_1.weight',
'h.10.mlp.c_fc.weight', 'h.9.ln_2.weight', 'h.8.mlp.c_proj.weight',
'h.8.attn.bias', 'h.2.ln_2.weight', 'h.2.ln_1.weight',
'h.5.attn.c_proj.bias', 'h.7.attn.bias', 'h.5.ln_2.bias',
'h.9.mlp.c_fc.bias', 'h.11.ln_1.weight', 'h.10.ln_2.bias',
'h.9.ln_2.bias', 'h.1.attn.c_attn.weight', 'h.11.attn.c_attn.weight',
'h.5.ln_1.weight', 'h.5.attn.c_attn.weight', 'h.5.mlp.c_proj.weight',
'h.1.attn.bias', 'h.11.ln_1.bias', 'h.0.ln_1.weight',
'h.7.mlp.c_proj.weight', 'h.7.mlp.c_proj.bias', 'h.7.ln_1.weight',
'h.0.attn.c_attn.weight', 'h.9.mlp.c_fc.weight', 'h.4.ln_1.bias',

```
'h.4.mlp.c_proj.bias', 'h.5.mlp.c_fc.weight', 'h.10.attn.c_proj.bias',
'h.8.ln_2.bias', 'h.6.attn.c_proj.bias', 'h.4.mlp.c_proj.weight',
'h.4.attn.c_attn.bias', 'h.10.ln_1.weight', 'h.3.mlp.c_proj.weight',
'h.10.attn.c_proj.weight', 'h.10.ln_2.weight', 'h.6.attn.bias',
'h.5.mlp.c_fc.bias', 'h.2.attn.c_attn.weight', 'h.5.ln_2.weight',
'h.5.attn.c_attn.bias', 'h.7.mlp.c_fc.weight', 'h.4.ln_1.weight',
'h.10.mlp.c_proj.bias', 'h.3.attn.c_attn.bias', 'h.3.ln_1.weight',
'h.8.attn.c_attn.weight', 'h.11.attn.c_attn.bias',
'h.0.attn.c_proj.weight', 'h.7.attn.c_proj.bias',
'h.3.mlp.c_proj.bias', 'h.11.mlp.c_proj.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

Tokenizer Length: 30527

/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:310
: FutureWarning: This implementation of AdamW is deprecated and will be
removed in a future version. Use the PyTorch implementation
torch.optim.AdamW instead, or set `no_deprecation_warning=True` to
disable this warning
  FutureWarning,

Optimizer: AdamW (
Parameter Group 0
    betas: (0.9, 0.999)
    correct_bias: True
    eps: 1e-06
    initial_lr: 0.001
    lr: 0.001
    weight_decay: 0.0
)

Build train and validation dataloaders

/usr/local/lib/python3.7/dist-packages/ignite/handlers/checkpoint.py:99
3: UserWarning: Argument save_interval is deprecated and should be
None. This argument will be removed in 0.5.0.Please, use events
filtering instead, e.g. Events.ITERATION_STARTED(every=1000)
  warnings.warn(msg)

Validation: {'average_nll': 2.5277858681611565,
 'average_ppl': 12.525741767371562,
 'nll': 2.5277858681611565}
Epoch [1/3]: 100% 267/267 [23:40<00:00,  5.34s/it, loss=0.0493,
lr=0.000729]

Validation: {'average_nll': 1.8213096610418915,
 'average_ppl': 6.179946787785855,
 'nll': 1.8213096610418915}
Epoch [2/3]: 100% 267/267 [23:58<00:00,  5.41s/it, loss=0.0326,
lr=0.000395]
```

```
Validation: {'average_nll': 1.7059233358677524,
 'average_ppl': 5.506467639743078,
 'nll': 1.7059233358677524}
Epoch [3/3]: 100% 267/267 [25:00<00:00,  5.64s/it, loss=0.0274,
lr=6.21e-5]

--------------------------------------------------------------------
```

Training on English dataset : *gpt2* (3 epochs)

```
Prepare tokenizer, pretrained model and optimizer - add special tokens
for fine-tuning

Model: <class 'transformers.models.gpt2.modeling_gpt2.GPT2LMHeadModel'>
Config: <class
'transformers.models.gpt2.configuration_gpt2.GPT2Config'>
Tokenizer: <class
'transformers.models.gpt2.tokenization_gpt2.GPT2Tokenizer'>

Using pretrained model

Downloading: 100% 1.04M/1.04M [00:00<00:00, 2.78MB/s]
Downloading: 100% 456k/456k [00:00<00:00, 1.46MB/s]
Downloading: 100% 665/665 [00:00<00:00, 389kB/s]
Downloading: 100% 548M/548M [00:10<00:00, 50.5MB/s]

Tokenizer Length: 50262

/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:310
: FutureWarning: This implementation of AdamW is deprecated and will be
removed in a future version. Use the PyTorch implementation
torch.optim.AdamW instead, or set `no_deprecation_warning=True` to
disable this warning
  FutureWarning,

Optimizer: AdamW (
Parameter Group 0
    betas: (0.9, 0.999)
    correct_bias: True
    eps: 1e-06
    initial_lr: 0.001
    lr: 0.001
    weight_decay: 0.0
)

Build train and validation dataloaders

/usr/local/lib/python3.7/dist-packages/ignite/handlers/checkpoint.py:99
3: UserWarning: Argument save_interval is deprecated and should be
```

```
None. This argument will be removed in 0.5.0.Please, use events
filtering instead, e.g. Events.ITERATION_STARTED(every=1000)
  warnings.warn(msg)

Validation: {'average_nll': 2.7318008973498054,
 'average_ppl': 15.36052485448808,
 'nll': 2.7318008973498054}
Epoch [1/3]: 100% 267/267 [55:40<00:00, 12.56s/it, loss=0.0932,
lr=0.000729]

Validation: {'average_nll': 1.5842364995078604,
 'average_ppl': 4.875567458932061,
 'nll': 1.5842364995078604}
Epoch [2/3]: 100% 267/267 [49:20<00:00, 11.13s/it, loss=0.0345,
lr=0.000395]

Validation: {'average_nll': 1.664853962183444,
 'average_ppl': 5.284901397941488,
 'nll': 1.664853962183444}
Epoch [3/3]: 100% 267/267 [49:22<00:00, 11.14s/it, loss=0.0276,
lr=6.21e-5]
```

Training on English dataset : *gpt2* (9 epochs)

```
Prepare tokenizer, pretrained model and optimizer - add special tokens
for fine-tuning

Model: <class 'transformers.models.gpt2.modeling_gpt2.GPT2LMHeadModel'>
Config: <class
'transformers.models.gpt2.configuration_gpt2.GPT2Config'>
Tokenizer: <class
'transformers.models.gpt2.tokenization_gpt2.GPT2Tokenizer'>

Using pretrained model


Tokenizer Length: 50262

/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:310
: FutureWarning: This implementation of AdamW is deprecated and will be
removed in a future version. Use the PyTorch implementation
torch.optim.AdamW instead, or set `no_deprecation_warning=True` to
disable this warning
  FutureWarning,

Optimizer: AdamW (
Parameter Group 0
    betas: (0.9, 0.999)
    correct_bias: True
```

```
    eps: 1e-06
    initial_lr: 0.001
    lr: 0.001
    weight_decay: 0.0
)

Build train and validation dataloaders

/usr/local/lib/python3.7/dist-packages/ignite/handlers/checkpoint.py:99
3: UserWarning: Argument save_interval is deprecated and should be
None. This argument will be removed in 0.5.0.Please, use events
filtering instead, e.g. Events.ITERATION_STARTED(every=1000)
  warnings.warn(msg)

Validation: {'average_nll': 3.0569875769255535,
 'average_ppl': 21.263406210126817,
 'nll': 3.0569875769255535}
Epoch [1/9]: 100% 267/267 [51:00<00:00, 11.51s/it, loss=0.0885,
lr=0.000919]

Validation: {'average_nll': 1.9028442619746295,
 'average_ppl': 6.704937947127965,
 'nll': 1.9028442619746295}
Epoch [2/9]: 100% 267/267 [42:36<00:00,  9.61s/it, loss=0.0361,
lr=0.000819]

Validation: {'average_nll': 1.424958921294465,
 'average_ppl': 4.157687046846091,
 'nll': 1.424958921294465}
Epoch [3/9]: 100% 267/267 [42:25<00:00,  9.57s/it, loss=0.0254,
lr=0.000719]

Validation: {'average_nll': 1.3377531989071743,
 'average_ppl': 3.8104725065153375,
 'nll': 1.3377531989071743}
Epoch [4/9]: 100% 267/267 [43:02<00:00,  9.71s/it, loss=0.0213,
lr=0.000619]

Validation: {'average_nll': 1.1823793194445722,
 'average_ppl': 3.262126617788971,
 'nll': 1.1823793194445722}
Epoch [5/9]: 100% 267/267 [42:45<00:00,  9.65s/it, loss=0.0192,
lr=0.000519]

Validation: {'average_nll': 0.9099034607300224,
 'average_ppl': 2.4840827102774146,
 'nll': 0.9099034607300224}
Epoch [6/9]: 100% 267/267 [42:36<00:00,  9.61s/it, loss=0.0173,
lr=0.000419]

Validation: {'average_nll': 0.7323501705240596,
```

```
 'average_ppl': 2.0799631360932067,
 'nll': 0.7323501705240596}
Epoch [7/9]: 100% 267/267 [43:53<00:00,  9.90s/it, loss=0.0125,
lr=0.000319]

Validation: {'average_nll': 0.6558546641722397,
 'average_ppl': 1.9267885710237618,
 'nll': 0.6558546641722397}
Epoch [8/9]: 100% 267/267 [43:20<00:00,  9.78s/it, loss=0.0114,
lr=0.000219]

Validation: {'average_nll': 0.5458546641722322,
 'average_ppl': 0.8267865710237768,
 'nll': 0.5058546676558773}
Epoch [9/9]: 100% 267/267 [31:54<00:00,  6.72s/it, loss=0.0101,
lr=0.000119]
```