A Configurations

We provide all model and training configurations used across our experiments:

A.1 BERT Experiments for Classification and Mask-Filling

- Model configuration for BERT-BASE classification and mask-filling:

```
attention_dropout_rate: 0.1
dropout_rate: 0.1
hidden_activation: gelu
hidden_size: 768
initializer_range: 0.02
intermediate_size: 3072
max_position_embeddings: 512
num_attention_heads: 12
num_layers: 12
type_vocab_size: 2
vocab_size: 30522
```

- Model configuration for BERT-LARGE classification and mask-filling:

```
attention_dropout_rate: 0.1
dropout_rate: 0.1
hidden_activation: gelu
hidden_size: 1024
initializer_range: 0.02
intermediate_size: 4096
max_position_embeddings: 512
num_attention_heads: 16
num_layers: 24
type_vocab_size: 2
vocab_size: 30522
```

- Training configuration for classification with BERT-BASE and in-domain data:

```
num_classes: 2
train_data:
  global_batch_size: 128
  seq_length: 512
validation_data:
  global_batch_size: 32
  seq_length: 512
trainer:
  max_to_keep: 3
  checkpoint_interval: 1000
  decay_steps: 30000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-5
  power: 1.0
  optimizer: adam
  warmup_steps: 5000
  steps_per_loop: 1000
  train_steps: 30000
  validation_steps: 256
```

- Training configuration for classification with BERT-LARGE and in-domain data:

```
num_classes: 2
train_data:
  global_batch_size: 128
  seq_length: 512
validation_data:
  global_batch_size: 32
  seq_length: 512
trainer:
  max_to_keep: 3
```

```
  checkpoint_interval: 1000
  decay_steps: 100000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-6
  power: 1.0
  optimizer: adam
  warmup_steps: 10000
  steps_per_loop: 1000
  train_steps: 100000
  validation_steps: 3000
```

- Training configuration for classification with BERT-BASE and out-domain data:

```
num_classes: 2
train_data:
  global_batch_size: 128
  seq_length: 512
validation_data:
  global_batch_size: 128
  seq_length: 512
trainer:
  max_to_keep: 3
  checkpoint_interval: 5000
  decay_steps: 500000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-6
  power: 1.0
  optimizer: adam
  warmup_steps: 10000
  steps_per_loop: 1000
  train_steps: 500000
  validation_steps: 512
```

- Training configuration for classification with BERT-LARGE and out-domain data:

```
num_classes: 2
train_data:
  global_batch_size: 128
  seq_length: 512
validation_data:
  global_batch_size: 128
  seq_length: 512
trainer:
  max_to_keep: 3
  checkpoint_interval: 5000
  decay_steps: 500000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-6
  power: 1.0
  optimizer: adam
  warmup_steps: 10000
  steps_per_loop: 1000
  train_steps: 500000
  validation_steps: 512
```

- Training configuration for mask-filling with BERT-BASE and in-domain data:

```
train_data:
  global_batch_size: 128
  seq_length: 512
  max_predictions_per_seq: 20
validation_data:
  global_batch_size: 128
  seq_length: 512
  max_predictions_per_seq: 20
trainer:
  checkpoint_interval: 2000
  max_to_keep: 30
```

```
  decay_steps: 30000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-8
  power: 1.0
  optimizer: adam
  warmup_steps: 5000
  steps_per_loop: 1000
  train_steps: 30000
  validation_interval: 1000
```

- Training configuration for mask-filling with BERT-LARGE and in-domain data:

```
train_data:
  global_batch_size: 128
  seq_length: 512
  max_predictions_per_seq: 20
validation_data:
  global_batch_size: 128
  seq_length: 512
  max_predictions_per_seq: 20
trainer:
  checkpoint_interval: 2000
  max_to_keep: 30
  decay_steps: 30000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-8
  power: 1.0
  optimizer: adam
  warmup_steps: 5000
  steps_per_loop: 1000
  train_steps: 30000
  validation_interval: 1000
```

- Training configuration for mask-filling with BERT-BASE and out-domain data:

```
train_data:
  global_batch_size: 512
  seq_length: 512
validation_data:
  global_batch_size: 512
  seq_length: 512
trainer:
  checkpoint_interval: 5000
  max_to_keep: 10
  decay_steps: 300000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-6
  power: 1.0
  optimizer: adam
  warmup_steps: 10000
  steps_per_loop: 1000
  train_steps: 300000
  validation_steps: 1000
```

- Training configuration for mask-filling with BERT-LARGE and out-domain data:

```
train_data:
  global_batch_size: 512
  seq_length: 512
validation_data:
  global_batch_size: 512
  seq_length: 512
trainer:
  checkpoint_interval: 5000
  max_to_keep: 10
  decay_steps: 300000
  end_learning_rate: 0.0
  initial_learning_rate: 1.0e-6
```

```
power: 1.0
optimizer: adam
warmup_steps: 10000
steps_per_loop: 1000
train_steps: 300000
validation_steps: 1000
```

A.2 T5 Experiments for Generation

- The training configuration for generation with T5-BASE and in-domain data:

```
encoder_seq_length: 512
decoder_max_length: 128
train_batch_size: 128
max_train_steps: 100000
valid_batch_size: 128
dropout_rate: 0.2
optimizer: adam
learning_rate: 1.0e-6
```

- The training configurations for generation with T5-BASE/LARGE and in-domain/out-domain data are similar as above, except that the `learning rate` is set to `5.0e-6` for T5-LARGE in-domain data, `5.0e-4` for T5-BASE out-domain data, and `1.0e-4` for T5-LARGE out-domain data.