System Benchmarks:

Total time: 111s

Intel Xeon VM

```
In [14]: %%timeit
         from transformers import MBartForConditionalGeneration, MBart50TokenizerFast
         import torch
         from intel_extension_for_pytorch.transformers.optimize import optimize

         1.01 µs ± 118 ns per loop (mean ± std. dev. of 7 runs, 1,000,000 loops each)

In [15]: %%timeit
         model = MBartForConditionalGeneration.from_pretrained("facebook/mbart-large-50-one-to-many-mmt")
         tokenizer = MBart50TokenizerFast.from_pretrained("facebook/mbart-large-50-one-to-many-mmt", src_lang="en_XX")

         2.18 s ± 638 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

In [16]: %%timeit
         article_en = "Morality or moral behaviour is not necessarily the result of philosophical reflections. Thesephilosophical reflecti
         model_inputs = tokenizer(article_en, return_tensors="pt")

         266 µs ± 1.93 µs per loop (mean ± std. dev. of 7 runs, 1,000 loops each)

In [17]: %%timeit
         # translate from English to Hindi
         generated_tokens = model.generate(
             **model_inputs,
             forced_bos_token_id=tokenizer.lang_code_to_id["hi_IN"]
         )

         12.9 s ± 737 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

In [18]: %%timeit
         translation = tokenizer.batch_decode(generated_tokens, skip_special_tokens=True)

         71.3 µs ± 13.2 µs per loop (mean ± std. dev. of 7 runs, 10,000 loops each)

In [20]: %%timeit
         translation

         8.18 ns ± 2.14 ns per loop (mean ± std. dev. of 7 runs, 100,000,000 loops each)
```

Time : 15s

Time Difference : 111 - 15 = 96s