

System Benchmarks:

✓
43s

[4] model = MBartForConditionalGeneration.from_pretrained("facebook/mbart-large-50-one-to-many-mmt")

pytorch_model.bin: 100% 2.44G/2.44G [00:38<00:00, 32.6MB/s]

/usr/local/lib/python3.10/dist-packages/torch/_utils.py:831: UserWarning: TypedStorage is deprecated
return self.fget.__get__(instance, owner)()

generation_config.json: 100% 261/261 [00:00<00:00, 14.8kB/s]

✓
4s

[5] tokenizer = MBart50TokenizerFast.from_pretrained("facebook/mbart-large-50-one-to-many-mmt", src_

tokenizer_config.json: 100% 528/528 [00:00<00:00, 28.0kB/s]

sentencepiece.bpe.model: 100% 5.07M/5.07M [00:00<00:00, 19.8MB/s]

special_tokens_map.json: 100% 717/717 [00:00<00:00, 42.0kB/s]

✓
2s

[7] article_en = "Morality or moral behaviour is not necessarily the result of philo

✓
2s

[8] model_inputs = tokenizer(article_en, return_tensors="pt")

✓
1m

[10]

translate from English to Hindi
generated_tokens = model.generate(
 **model_inputs,
 forced_bos_token_id=tokenizer.lang_code_to_id["hi_IN"]
)

✓
0s

[11] translation = tokenizer.batch_decode(generated_tokens, skip_special_tokens=True)

✓
0s

translation

['नैतिकता या नैतिक व्यवहार अनिवार्य रूप से दार्शनिक प्रतिबिम्बों का परिणाम नहीं है, इन दार्शनिक प्रतिबिम्बों को तार्किकीकरण की भावना की आवश्यकता होती है क्योंकि कभी-कभी हमें अपने निर्णयों, अपने अनुमानों, हमारी धारणाओं और निश्चित रूप से अपने कार्यों को न्यायसंगत करना होता है, मानव समाजी और संस्कृतियों की नैतिक धारणाएँ, रीति-रिवाजों, विश्वासों और अभ्यासों का एक दिन के भीतर अस्तित्व नहीं हुआ था; बल्कि वे समय के दौरान अपने सामाजिक या सांस्कृतिक विकास की प्रक्रिया में विभिन्न स्थितियों के परिणामस्वरूप विकसित हुए थे, और एक प्रतिबिम्ब प्रक्रिया के रूप में नैतिकता इन नैतिकताओं पर प्रतिबिम्ब करती है, यही कारण है कि इसे नैतिक दर्शन भी कहा जाता है, यह दार्शनिक प्रतिबिम्बों से संबंधित है और हम जो नैतिक निर्णय करते हैं उनका विश्लेषण करता है, हम']

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. These philosophical 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