New issue

Pre-train Data Structure #37

Closed

saifulislam79 opened last month · edited by saifulislam79 Edits

Thank you for sharing great work, I want to know about pre-train data format and it meaning given config file

```
> `
> # Datasets
> text_QA_dataset: <speech input-ids>
> TTS_dataset: <text-input-ids>
```

Basically i want to know how can i prepare `text_QA_dataset` and `TTS_dataset` and it's format structure. i am waiting for your response and great-full to you.

What is the different between `text_QA_dataset` and `TTS_dataset` .

amuvarma13 last month · edited by amuvarma13 Edits Collaborator

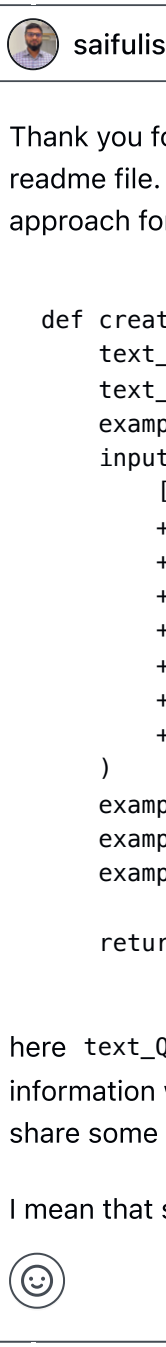
```python
tokeniser_length = 128256
start_of_text = 128000
end_of_text = 128009

start_of_speech = tokeniser_length + 1
end_of_speech = tokeniser_length + 2

start_of_human = tokeniser_length + 3
end_of_human = tokeniser_length + 4

start_of_ai = tokeniser_length + 5
end_of_ai =  tokeniser_length + 6
pad_token = tokeniser_length + 7

audio_tokens_start = tokeniser_length + 10
```

start of human --- start of text --- text tokens --- end of text--- end of human--- start of ai --- start of speech --- speech tokens --- end of speech --- end of ai

Let me know if unclear or further questions.

EDIT - for text which I realise you also asked about:

start of human --- start of text --- question text tokens --- end of text--- end of human --- start of ai --- start of text --- answer text tokens --- end of text --- end of ai



saifulislam79 last month · edited by saifulislam79 Edits Author

Thank you for your reply i had reviewed data processing code into colab, which mentioned into readme file. I need more clear understanding the processing approach, Is it same processing approach for both fine-tune and pre-train .

```
def create_input_ids(example):
    text_ids = tokenizer.encode(example["text"],  add_special_tokens=True)
    text_ids.append(end_of_text)
    example["text_tokens"] = text_ids
    input_ids = (
        [start_of_human]
        + example["text_tokens"]
        + [end_of_human]
        + [start_of_ai]
        + [start_of_speech]
        + example["codes_list"]
        + [end_of_speech]
        + [end_of_ai]
    )
    example["input_ids"] = input_ids
    example["labels"] = input_ids
    example["attention_mask"] = [1] * len(input_ids)

    return example
```

here `text_QA_dataset` and `TTS_dataset` why mentions separately. `text_QA_dataset` is QA textual information with audio or `TTS_dataset` is as normal TTS dataset. it will more convenient , if possible share some data sample about `text_QA_dataset` and `TTS_dataset` format.

I mean that same format as like fine-tune dataset but use different dataset or other.

amuvarma13 last month · edited by amuvarma13 Edits Collaborator

Yep the `text_QA_dataset` is only text no audio. `tts_dataset` is text and then a spoken version of the text.

Here is what a text sample could look like, all the text samples are chained together so all input_ids are the same length (8192) for pretraining to make the training as efficient as possible:

start of human --- start of text --- question text tokens (i.e. AutoTokeniser.tokenise("What is 2 +2?") --- end of text--- end of human --- start of ai --- start of text --- (i.e. AutoTokeniser.tokenise("Great question, 2 + 2 =4") --- end of text --- end of ai

👍 1 👀 1

amuvarma13 last month Collaborator

Feel free to close this issue - if your question is answered!

❤️ 1

saifulislam79 last month · edited by saifulislam79 Edits Author

This is the last clarification
Example with Token IDs (simplified illustration)
Assume the tokenizer produces the following (again, just for illustration):
input sentence 1 : What is 2 + 2? ----> audio1.mp3

Answer other sentence : Great question, 2 + 2 = 4. ---> audio2.mp3

```
"start of human" → [101]
"start of text" → [102]
"What is 2 + 2?" → [2001, 2002, 2003, 2004, 2005]
"end of text" → [103]
"end of human" → [104]
"start of ai" → [105]
"start of text" → [102]
"Great question, 2 + 2 = 4." → [3001, 3002, 3003, 3004, 3005, 3006]
"end of text" → [103]
"end of ai" → [106]
```

Chained together example of question and answer:

```
[101, 102, 2001, 2002, 2003, 2004, 2005, 103, 104, 105, 102, 3001, 3002, 3003, 3004, 3005, 3006, 103, 106]
```

if i have 1M text sentences and it's corresponding audio codes, what will be `<speech input-ids>` and `<text-input-ids>` . Could you please give a example .

saiful9379 last month

@amuvarma13 thank for your clarification.

amuvarma13 last month Collaborator

Sure,
Text input ids (text dataset) is for text question text answer pairs - the format you have given above is correct.
Speech input ids i.e. the tts dataset is for text speech pairs no question answering - the format I gave above with start of speech etc is what you want for this,.

1

amuvarma13 last month Collaborator

Marking as solved - reopen if unclear.

amuvarma13 closed this as completed last month

amuvarma13 mentioned this 3 weeks ago
Pre-training datasets samples #20

amuvarma13 mentioned this 2 weeks ago
Datasets for build model pretrained #80

Sundragon1993 3 days ago

@amuvarma13 @saiful9379 Would it be possible to share the preparation code for the text_QA_dataset, perhaps by adapting the TTS preparation code?

Add a comment

Write Preview

Use Markdown to format your comment

Paste, drop, or click to add files

Reopen Issue

Comment

Metadata

Assignees

No one assigned

Labels

No labels

Type

No type

Projects

No projects

Milestone

No milestone

Relationships

None yet

Development

Code with Copilot Agent Mode

No branches or pull requests

Notifications Customize

Unsubscribe

You're receiving notifications because you're subscribed to this thread.

Participants