Topics

My posts

More

CATEGORIES

General

Site Feedback

All categories

Skip to main content

2025 USA-NA-AIO Round 2, Problem 3, Part 10

USAAIO May 2025

Part 10 (5 points, non-coding task)

In this part, you are asked to answer some questions about a CLIP model that you shall build in the next part.

- Write your answers in the text cell below.
- To get answers, you may need to run experimental code to better learn the ViT and Bert models.
- We only grade your answers in the text cell.

1. **Image encoder**
   - Define `model_image = ViTModel.from_pretrained('google/vit-base-patch16-224')`. We use all blocks except the last pooler layer. That is, this ViT model has two outputs: with their key names as `last_hidden_state` and `pooler_output`. You should take the value associated with the key `last_hidden_state`.
   - From the last hidden state, we project from position 0 to a latent space with dimension `embedding_size` (e.g., 512). The output is called image embedding.
2. **Text encoder**
   - Define `model_text = BertModel.from_pretrained('bert-base-uncased')`. We use all blocks except the last pooler layer. That is, this Bert model has two outputs: with their key names as `last_hidden_state` and `pooler_output`. You should take the value associated with the key `last_hidden_state`.

  - From the last hidden state, we project from position 0 to a latent space with dimension `embedding_size` (e.g., 512). The output is called text embedding.

Answer the following questions. (Reasoning is required only for Question 3)

1. Let `image_batch` be with shape `(B,3,224,224)`. What is the shape of `model_image(image_batch)`[last_hidden_state] ?

2. Let `token_id_batch` and `attention_mask_batch` be with shape `(B,L)`. What is the shape of `model_text(input_ids = token_id_batch, attention_mask = attention_mask_batch)['last_hidden_state']` ?

3. For both the image encoder and the text encoder, we project the last hidden state from position 0 to a latent space with the same dimension `embedding_size`.

   3.1. Why do we add this additional out-projection layer?

   3.2. Why this layer is added on position 0 only?

   3.3. Why the output dimensions from these two encoders are the same?

USAAIO **May 2025**

```
### DO YOU EXPERIMENTAL STUDY HERE ###

image_batch, token_id_batch, attention_mask_batch = next(iter(CLIP_dataloader))

model_image = ViTModel.from_pretrained('google/vit-base-patch16-224')

print(model_image(image_batch).keys())

print(model_image(image_batch)['last_hidden_state'].shape)

model_text = BertModel.from_pretrained('bert-base-uncased')
```

Skip to main content

```
print(model_text(input_ids = token_id_batch, attention_mask = attention_mask_batc

print(model_text(input_ids = token_id_batch, attention_mask = attention_mask_batc

""" END OF THIS PART """
```

Misplaced '#'

1. `(B,197,768)`.

2. `(B,L,768)`.

3.

   3.1. Ensure that the embedding of an image and the embedding of a text are similar to each

   3.2. In both ViT and BERT, position 0 in the last hidden state stores information of the whole image and text, respectively.

   3.3. We need to compute their similarity, such as the cosine similarity. So this requires their dimensions to be same.

""" END OF THIS PART """

Skip to main content

Related topics

Topic	Replies	Activity
2025 USA-NA-AIO Round 2, Problem 3, Part 11	1	May 2025
2025 USA-NA-AIO Round 2, Problem 3, Part 13	1	May 2025
2025 USA-NA-AIO Round 2, Problem 3, Part 5	1	May 2025
2025 USA-NA-AIO Round 2, Problem 3, Part 1	1	May 2025
2025 USA-NA-AIO Round 2, Problem 3, Part 15	1	May 2025