Persona Chatbot

Hsuan Su 2021.12.19

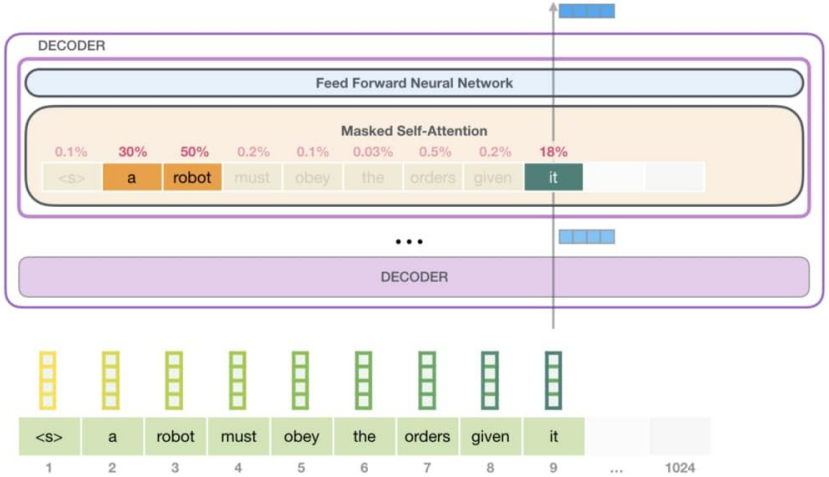
Outline

- Chatbot Recall
- Persona Data
- Persona Chatbot
- Requirements
- Source

Chatbot Recall

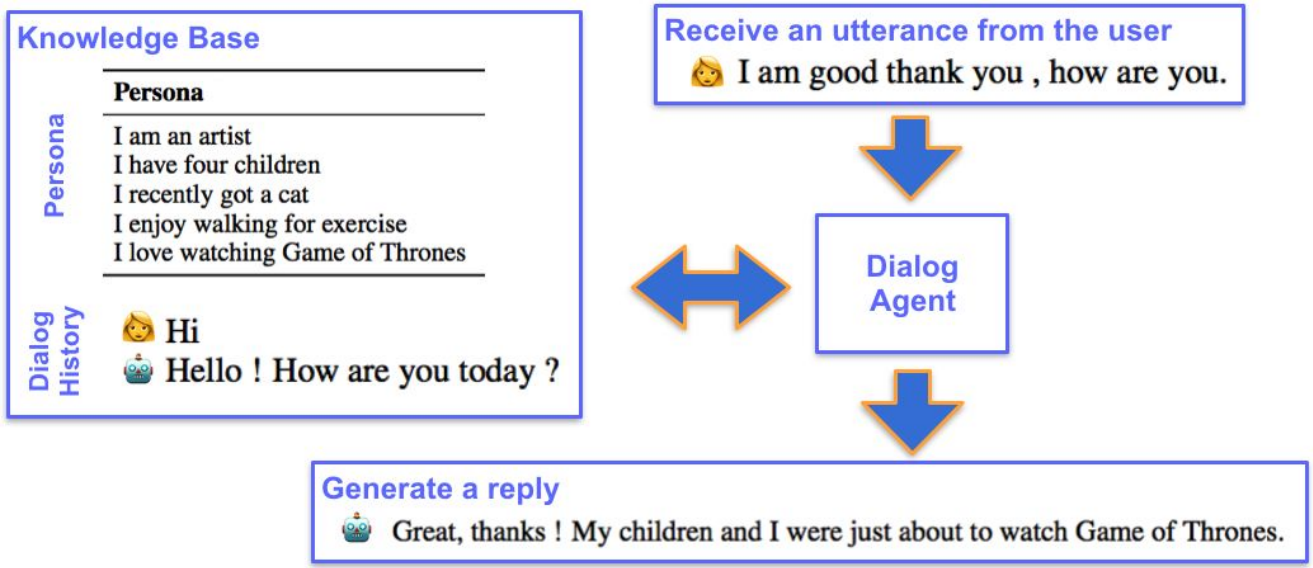
- RNN-based Seq2seq Chatbot https://github.com/Conchylicultor/DeepQA
- Mostly fine-tuned pretrained GPT-2 LM on Dialogue corpus

  .https://github.com/microsoft/DialoGPT

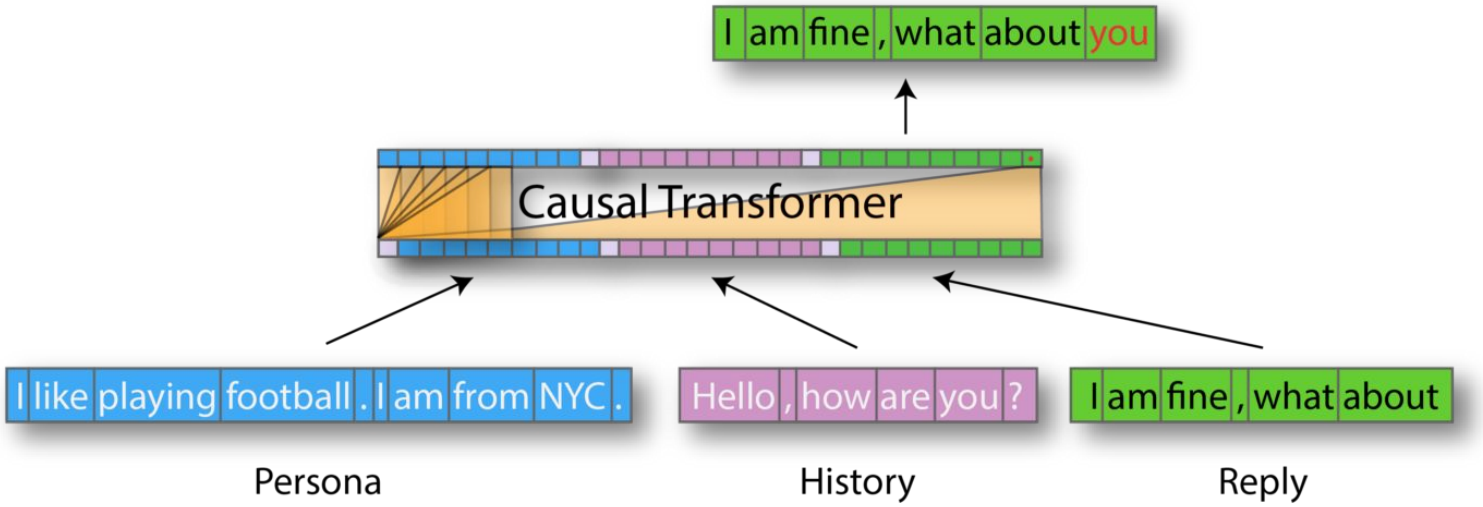


Persona Chatbot

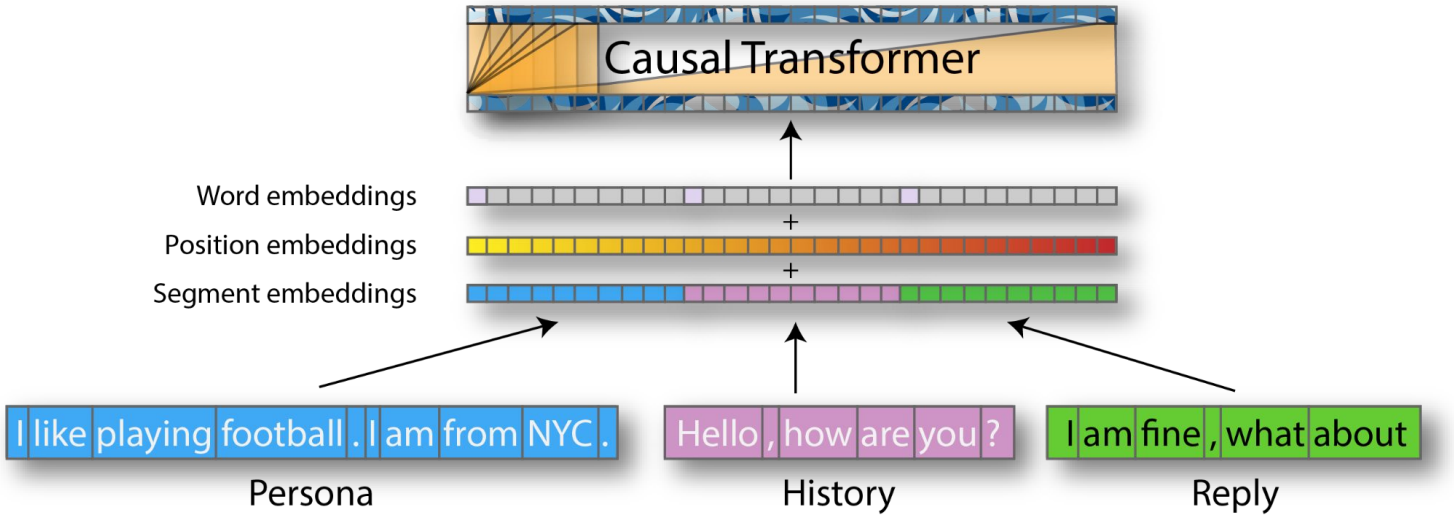
Persona Dataset



Model Input

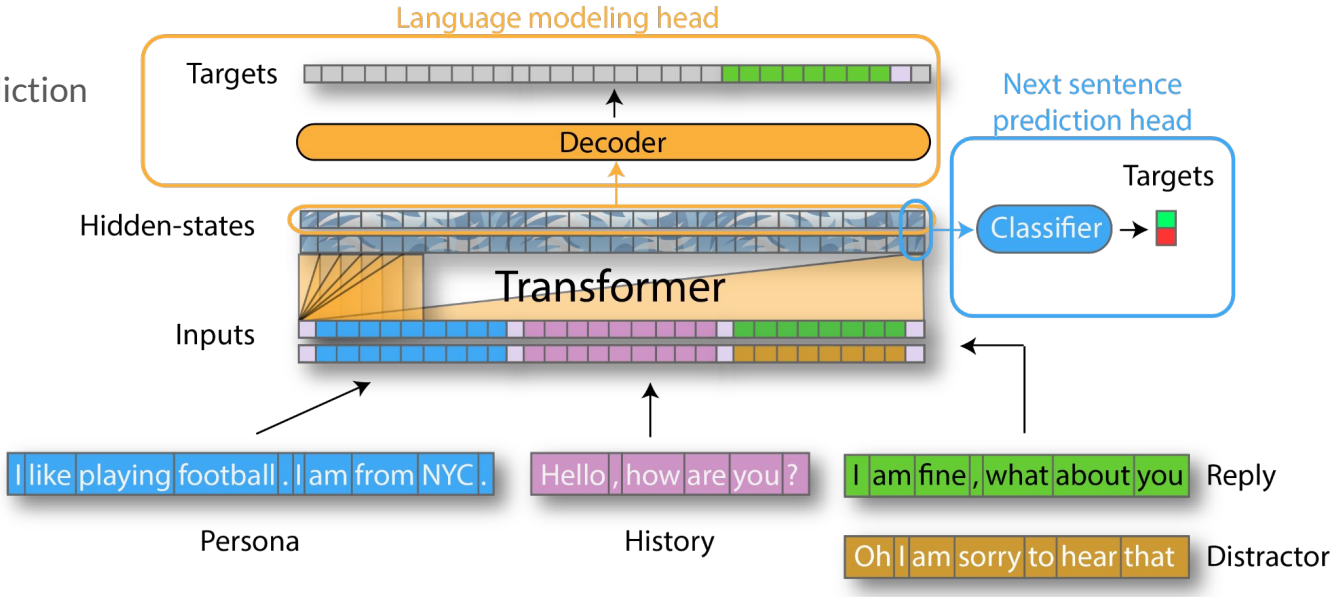

I am fine , what about you
Causal Transformer
I like playing football . I am from NYC .
Hello , how are you ?
I am fine , what about
Persona
History
Reply

Model Input


Causal Transformer
Word embeddings
Position embeddings
Segment embeddings
I like playing football . I am from NYC .
Hello , how are you ?
I am fine , what about
Persona
History
Reply

Multipule Losses

- Language modeling
- Next-sentence prediction



```
parser.add_argument("--model_checkpoint", type=str, default="openai-gpt", help="Path, url or short name of the model")
```

Requirements

- Trained a Persona bot by yourself with **GPT-2** model
- Try other pretrained GPT-2 model e.g. **DialoGPT(microsoft/DialoGPT-medium)**
- Try to interact with it :)
- Have Fun

The training script can be used in single GPU or multi GPU settings:

```
python ./train.py  # Single GPU training
python -m torch.distributed.launch --nproc_per_node=8 ./train.py  # Training on 8 GPUs
```

The training script accept several arguments to tweak the training:

Argument	Type	Default value	Description
dataset_path	`str`	`""`	Path or url of the dataset. If empty download from S3.
dataset_cache	`str`	`'./dataset_cache.bin'`	Path or url of the dataset cache
model	`str`	`"openai-gpt"`	Path, url or short name of the model
num_candidates	`int`	`2`	Number of candidates for training
max_history	`int`	`2`	Number of previous exchanges to keep in history
train_batch_size	`int`	`4`	Batch size for training
valid_batch_size	`int`	`4`	Batch size for validation
gradient_accumulation_steps	`int`	`8`	Accumulate gradients on several steps
lr	`float`	`6.25e-5`	Learning rate
lm_coef	`float`	`1.0`	LM loss coefficient
mc_coef	`float`	`1.0`	Multiple-choice loss coefficient
max_norm	`float`	`1.0`	Clipping gradient norm
n_epochs	`int`	`3`	Number of training epochs
personality_permutations	`int`	`1`	Number of permutations of personality sentences

Source

Persona Chat Dataset: https://gist.github.com/thomwolf/ecc52ea728d29c9724320b38619bd6a6

Persona Bot 好讀版:
https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313

Colab:
https://colab.research.google.com/drive/16-y2BToRAwwNpQLMBhfgLh2RozzkvsV5?usp=sharing

Persona Bot Repo:
https://github.com/huggingface/transfer-learning-conv-ai