A neural network based vietnamese chatbot

Business, Management

Nowadays, chatbot is a hot topic, chatbots are built from generative models are gaining success. The purpose of this article is to build a Vietnamese chatbot based on the seq2seq model incorporating the attention mechanism. We have built the model and tested on deep learning framework Pytorch using GPU. The model was trained end-to-end with no hand-crafted rules. Model is built from a small dataset and can generate responses to a user. However, generated responses still need to be improved to get a meaningful conversation.

Instroduction

Conversational agents or chatbots have been growing in popularity in recent years. Chatbot is a hot topic and chatbots are increasingly used widely in the modern society such as chatbot helps to book a hotel room, airplane ticket or to buy products from a shop. Choosing the model to build a chatbot will depend on the purpose and the space of the application. There are two types of chatbot models: selected model and generative model. The selected model is the model in which the responses were predefined, the answers are usually grammatical and the pattern is suitable for use in the service industry such as hotel reservations, ticket reservations, etc. But there is a problem that this model does not handle is that it can’t handle the questions that don’t have a predefined response. Due to the advances in the field of deep neural networks, chatbots become a hot topic, especially in natural language processing (NLP). Generative models are built to solve the problem of responding to non-predefined response. Generative models can handle new cases because they do not rely on any predefined response and create their own response starting from the person they must respond to. These

models are special they can give the users the feeling of talking to a real human. Since there are no predefined answers, these models need to learn how to build responses using a large collection of conversations.

One of the top models for building a generative chatbot is the sequence to sequence model. Recent the studies on sequence to sequence network with the attention mechanism have been achieved great successes. And this model has been successfully used for many different natural language processing tasks, such as alignment, translation, and summarization.

Conversational modeling can be phrased as a mapping between utterances and responses, and therefore can benefit from the encoder-decoder setup. That’s why we realized the success of building a chatbot using seq2seq with attention. We test the model on chat sessions from Vietnamese dataset of conversations, and find that the model can response to the users utterance. So, we believe that our model is sometimes able to produce natural conversations.

Related Work

Our approach inspired by the recent success of sequence to sequence learning with neural network on English-French translation task from the WMT’14 dataset which uses a multilayered Long Short-Term Memory (LSTM) to map sequence to sequences. It has also been used for other task of neural dialog models was proposed by Vinyals and Le, where a dialog response is generated from a dialog question in a sequence to sequence framework.

Our approach is based on recent work which proposed to translate based on Sequence to Sequence Network and Attention. This model used for neural machine translation and achieves improvements on the English-French translation task from the question on Open Data Stack Exchange. Recently, in the paper of neural machine translation by jointly learning to align and translate, Bahdanau et al. (2015) has successfully applied such attention mechanism to jointly translate and align words.

Building bots and conversational systems has been pursued by many researchers over the last decades, there are a number of articles on chatbot building methods that we do not list here. Our project focused on building Vietnamese chatbot model based on seq2seq network and attention mechanism. And we have built our model on a different framework (Pytorch) to minimize the time of the training.

Model

We use a model very closely based on Pytorch’s Neural Machine Translation model by Sean Robertson, 2017. It’s a sequence-to-sequence model with attention mechanism that allows decoder more direct access to hidden state output by the encoder. The Seq2seq network is the model consisting of two recurrent neural networks (RNN). The encoder is the utterance/question by human that outputs a single vector, and the decoder reads that vector and output response. To the best of our knowledge, if only the context vector is passed between the encoder and decoder, that single vector carries the burden of encoding the entire sentence. So, attention mechanism introduced by Bahdanau et al., 2015 is exactly to resolve this problem. With an

attention mechanism, the input sentence is divided into n parts and it allows the decoder to focus on the relevant part of input. So, for every step the decoder can uses specific parts of the encoder’s outputs. Attention is calculated with another feedforward layer in the decoder. This layer will use the current input and hidden state to create a new vector, which is the same size as the input sequence. This vector is processed through softmax to create attention weights, which are multiplied by the encoders’ outputs to create a new context vector, which is then used to predict the next output..

Datatset

Data

The dataset of this paper is based on different sources of data. The data in Vietnamese was collected from the websites of learning English for Vietnamese that include the conversations that have been translated into Vietnamese. The dataset is good because it is based on conversations that is used to teach language and it was exactly translated into Vietnamese. So the data isn’t noisy and does not include offensive words such as those taken from social networks. Vietnamese dataset includes 1331 pairs of questions/utterance and responses. Sample of data is a line which include a pair of utterance and response. The file is a slash separated list of pairs.

The full process for preparing the data is:

- Read text file and split into lines, split lines into pairs
- Normalize text, filter by length and content
- Make word lists from sentences in pairs: vocabulary of utterances and vocabulary of responses.

Vietnamese word segmentation

Word segmentation is an important step in NLP, especially for the languages of East Asia in a single language type, example: Chinese Japanese, Thai, and Vietnamese. In languages of this type, the word boundary is not simply whitespace, as English word, but there is a strong connection between the single words, a compound word can be composed of one or more single words. Because of the above reasons, before implementing the training model, we have extracted the Vietnamese words and used Python Vietnamese Toolkit (pyvi) with the accuracy to 97, 86%.

Experiments

We pretrained the model on conversational datasets for 300, 000 steps of 1331 pairs of sentences. After filtering and selecting sentences of 15 words in length (including punctuation), collected 887 pairs of utterances and response. The vocabulary from utterances includes 843 words and the vocabulary from responses includes 918 words. We used Google Colab (free GPU) for training the model, the training time is 288 minutes and 1 second. Two examples of talking to the chatbot about reading and time are shown in Fig. 2 and Fig. 3.

The quality of the model was evaluated by measuring the perplexity of its responses compared to a separate validation dataset. Perplexity of the whole training process is 6, 845 and BLEU score of the whole model is 1. 443 %. Below in Fig. 4 shown a plot of training loss that was created while training.

A useful property of the attention mechanism is its highly interpretable outputs. Because it is used to weight specific encoder outputs of the input

sequence, we imagined looking where the network is focused most at each time step. The Fig. 5 describes the interpretation of the input sentence as “ cháu học lớp mấy” and the output sentence is: “ cháu học lớp 6”.

In Fig. 6 each pixel shows the weight wij of the annotation of the j-th source word for the i-th target word, in cale (0: light pink, 1: dark blue).

Conclusions

In this article, we have implemented a simple language model based on the seq2seq framework with attention decoder. Our modest results show that it can generate simple and basic conversations and extract the knowledge from generic dataset. The model has obvious limitations as the training size is limited and the answer is general but the model also has advantages such as the model built on Pytorch simple compact and significantly reduced time when training. In the future we plan to try larger dataset over multiple GPUs to train, and might switch to retrieval based chatbots.