Hierarchical Attention Network (HAN) model for depression detection

What is HAN?

"Hierarchical Attention Network for Document Classification"
<https://www.cs.cmu.edu/~./hovy/papers/16HLT-hierarchical-attention-networks.pdf>

HAN is a neural network architecture designed to perform text classification on documents. It is unique in that it is designed to mimic the hierarchical structure of documents. Similar to how documents can be decomposed into sentences and then further into words, this model operates on both the word and sentence level, incorporating two attention-based mechanisms for each level. This allows the model to identify important content for classification at different levels of the document hierarchy.

How-to guide

For this model, I highly recommend using Google Colaboratory to run experiments. Because of the large number of parameters in Transformer models, and relatively long length of each transcript, it is quite easy to run into memory allocation errors when performing backward pass on the loss. Colab provides remote access to systems with large memory resources that can help to mitigate this issue.

To use this model on Colab, download the notebook named "depression-HAN-daicwoz.ipynb" and upload it to your Google Drive account.

1. Clone the repository

If you are cloning locally, you only need to use this command.

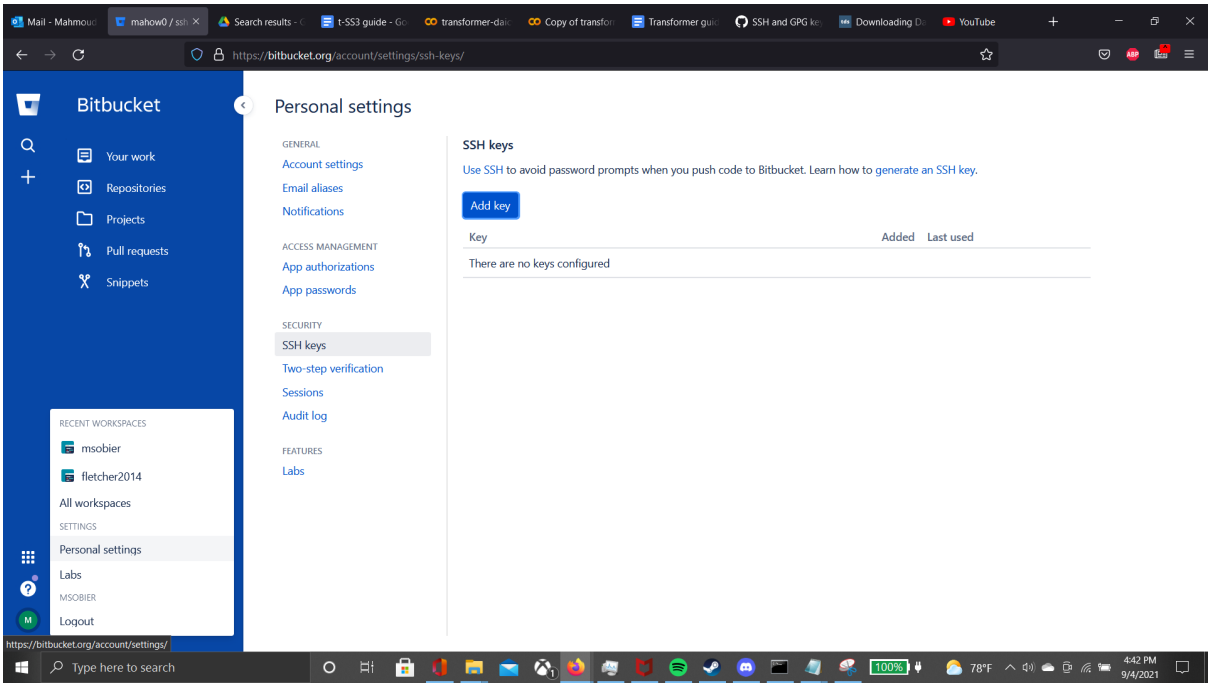
git clone https://[your bitbucket username]@bitbucket.org/fletcher2014/depression_model_attention_network

On Google Colab, the process is a bit more involved. You will need to generate and link an SSH key to verify your identity.

The first code block generates an SSH key. It will prompt you three times for keyboard input but you can leave each text field blank.

The second one will output the public key which you will need to give to Bitbucket in order for git to correctly identify you and clone the repository.

Once you have the public key, attach it to your Bitbucket account. To do this, click on your account icon on the bottom left and select “Personal settings.” Then, go to the tab labeled “SSH keys” under “Security.” Finally, click on “Add key” and paste the public key in the window that pops up.



After this step, you should have the ability to clone the repository. Note that you will need to do this step everytime you restart the Colaboratory runtime.

2. Dependencies

gensim==4.0.1
tqdm==4.62.0
torch==1.9.0+cu111
transformers==4.8.2
pytorch_ignite==0.4.6
imbalanced_learn==0.4.3
numpy==1.19.5
icecream==2.1.1
pandas==1.1.5
ignite==1.1.0
imblearn==0.0

If you're not sure whether you have any of these packages, install them by going into the root folder of the repository and using the command:

pip install -r requirements.txt

3. Configure the model

You can configure the model hyperparameters by manipulating the dictionary called "model_config." The fields "embed_dim", "hidden_dim", "attn_dim" determine the dimensions of the embeddings, hidden state, and attention mechanisms respectively. "Num_layers" specifies the number of stacked GRU layers and "dropout" is the dropout probability. You can adjust the learning rate of the model by changing the value of the 'lr' field in "optim_params."

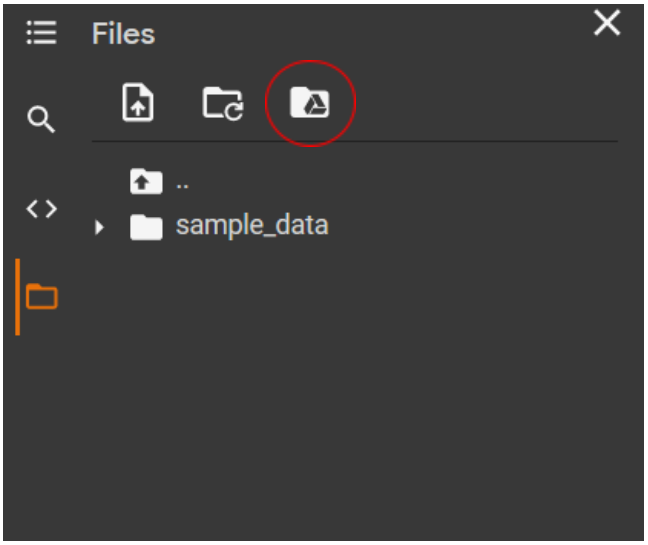
Training and evaluation parameters are specified by the dictionaries "train_set_params" and "eval_set_params" respectively, found in main.py. In a Google Colaboratory environment, I recommend setting the batch size to some value between 8 and 16. Any higher than that and it is likely that you will run out of memory on the GPU.

4. Run training and evaluation loop

To begin training the model, place the .csv files specifying the three DAICWOZ dataset splits (train, validation, evaluation) in the same folder and all the transcript files in another. Run main.py, located in src/, with the following command:

python main.py --train_csv_path [path to training set csv] --val_csv_path [path to validation set csv] --eval_csv_path [path to evaluation set csv] --lr [initial learning rate] --num_epochs [number of epochs]

For Colab, you will need to have a copy of the dataset linked to your current runtime. I recommend mounting your Google Drive file system, which you can do by accessing the folder tab on the right hand side and clicking on the button that looks like a folder with the Google Drive logo on it.



Save the contents of the “chatbot-depression-DAIC-WOZ-database” Dropbox folder in your Google Drive account and now the dataset should be accessible from Google Colab.