Generate Answer to Visual Questions in the Medical Domain: Phase 1

Niki Nezakati - 98522094

1 Gathering data

The source of data is SLAKE 1.0 dataset, A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering [ISBI 2021 oral] [1]. This dataset contains 642 images and 14K QA pairs When you run the gathering data script, the program downloads this dataset from its source website. To do this, you should run the bash script located in the run folder named gather_raw_data.bash. You can add one argument to this bash which is the output directory and you can specify that or leave it as the default value (data/raw).

```
bash run/gather_raw_data.bash
```

After executing the provided bash script, you will have access to the resulting dataset. This dataset contains medical images along with corresponding question-answer pairs in both English and Chinese. In the downloaded folder, you will find subfolders containing the images, disease types presented in CSV format, mask file, as well as JSON files for training, validation, and test data. Here is the the amount of instances for each subfolder:

Raw Train Data	Raw Validation Data	Raw Test Data
9835	2099	2094

2 Preprocessing

To run preprocessing on the data, you should run the bash script located in run folder with name preprocess_raw_data.bash. You can add one argument to

this bash which is the output directory and you can specify that or leave it as the default value (data/preprocessed).

```
bash run/preprocess_raw_data.bash
```

This preprocessing step selects the English data which will be used in the project. This is done by selecting dictionary instances that have the key-value pair of ['q_lang'] == 'en' in the JSON files for training, validation, and test data. The resulted data is saved as JSON files for training (4919 instances), validation (1053 instances), and test (1053 instances) data. Here is the the amount of instances for each subfolder:

English Train Data	English Validation Data	English Test Data
4919	1053	1061

3 Tokenization

To run tokenization on the English data, you should run the bash script located in run folder with name break_data.bash. You can add two arguments to this bash which are the output directory for word and sentence tokenization. You can specify that or leave it as the default value (data/wordbroken) and (data/sentencebroken).

```
bash run/break_data.bash
```

This step involves tokenizing the English questions and answers into individual words and sentences. Word tokenization is done by splitting the English questions and answers by punctuation marks. In the sentence tokenization process, we include each question individually since they consist of only one sentence. The resulted data is saved as JSON files for training, validation, and test data. Here is the the amount of instances for each subfolder:

Train Question Words	Validations Question Words	Test Question Words
39573	8868	8722

Train Question Sents.	Validations Question Sents.	Test Question Sents.
4919	1053	1061

Train Answer Words	Validations Answer Words	Test Answer Words
6912	1537	1558

4 Statistics

To run Statistics on the English data, you should run the bash script located in run folder with name stats.bash.

```
bash run/stats.bash
```

This step involves finding unique answer (words), and questions (sentences) in train, validation, and test labels. We categorize them by the common ones (unique words that appear in more than one label), or uncommon ones. The code contains a count of word occurrence. The resulted data is saved as CSV files for training, validation, and test data. Here is the the amount of instances for each subfolder:

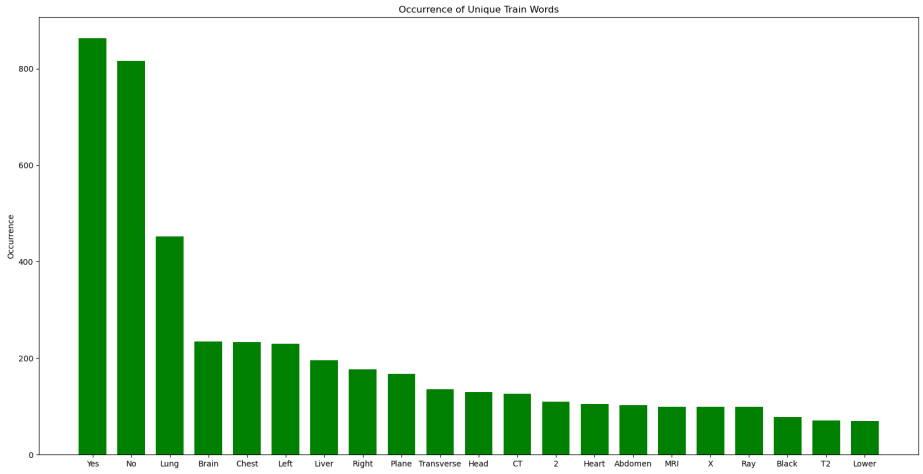
Unique Train Question	Unique Validation Question	Unique Test Question
284	245	252

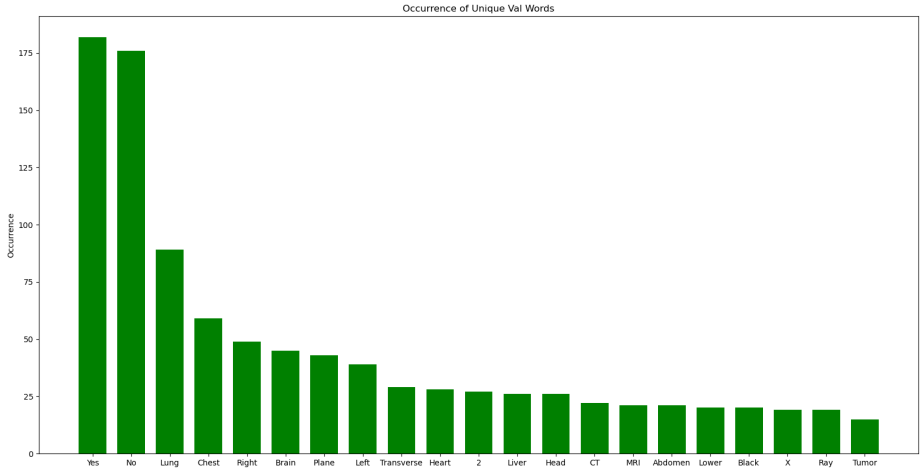
Unique Train Answers	Unique Validation Answers	Unique Test Answers
275	191	189

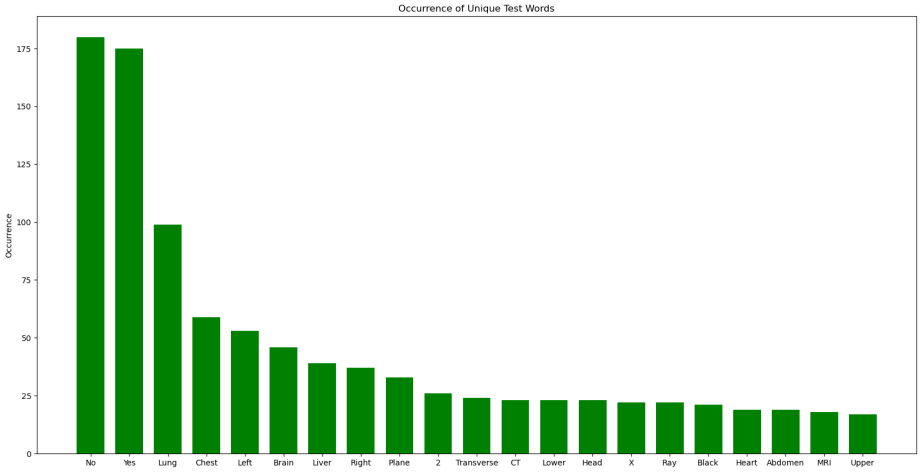
Common Uniq. Train Ans.	Common Uniq. Valid Ans.	Common Uniq. Test Ans.
230	137	139

Uncom. Uniq. Train Ans.	Uncom. Uniq. Valid Ans.	Uncom. Uniq. Test Ans.
45	54	50

Finally, the histogram displayed below represents the frequency of occurrence for the top 20 most unique label tokens, considering the large number of distinct tokens involved in the answers.







Note that due to the nature of the labels being comprised of single or multiple distinct words, it is not meaningful to assess measures such as the frequency of occurrence of label tokens, or identifying the most commonly appearing words.

References

[1] SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, Xiao-Ming Wu (2021). https://arxiv.org/abs/2102.09542