Proposal for the Final Project in the Course Computer Vision and Deep Learning SoSe 23

Does the Vision-Based Method Improve the Robustness of Noised Text Recognition?

Team

Han Yang (12531717):	Master's student in Computer Science / Computer Linguistics
Yian Yu (12676653):	Master's student in Computer Science
Jiawen Wang (12218743):	Master's student in Computer Linguistics / Computer Science

Problem Definition

The idea of this proposal comes from the paper "Robust Open-Vocabulary Translation from Visual Text Representations"[1] published on EMNLP in 2021. This work focused on the robustness of the text translation, when some letters are substituted. For example, the word pairs, *apple* and *ɑpple*, where the letter *a* is replaced by the Greek alphabet *alpha*, or *Apple* and *4pple*, where the letter *A* is replaced by the number *4*, are very easy for human beings to notice, be distinguished and be identified as the same word. Such letter substitutions are very common on the web community like Reddit. For the text-based machine translation techniques however, the substitution brings additional difficulty for machines to identify such word pairs as the same word, and it will even cause the out-of-vocabulary (OOV) problem because the substituted words like *4pple* do not exist in the dictionary at all.

In this work, the authors implemented a pipeline, where they first transfer the text-based data into images by using *pygame*[2] package with specific font and font size. Then the generated image will be used as the input of a CNN model, and the output of the following linear layer afterward will be considered as a representation of the input sentence. The produced representation will be used as the input of a standard transformer (including an encoder and a decoder), which is the component responsible for the machine translation task. During the training, the transformer is frozen, and the loss is calculated with the corresponding translation in the target language. In other words, the authors used the image-based method to obtain the representation of the text-based data, which may contain noise, to improve the robustness of the machine translation, inspired by the human vision robustness.

The authors have done exhaustive experiments, including the influence of various hyperparameters like the stride, and filter size of the CNNs, and evaluated the robustness against various noises. However, the author only used one font for rendering images, namely the *Google Noto font*. Though this is a pity for this work, it is an exciting research gap. Inspired by this point, in this proposal, we would like to explore the influence of the font

[1] https://aclanthology.org/2021.emnlp-main.576/

[2] https://www.pygame.org/news

on robustness. In the original task, machine translation, the transformer was used as the component for machine translation, but it might bring additional complexity and difficulty for a final project. We suggest using *Noised Text Recognition* as the task and thus discarding the transformer part. This project will use a vision-based approach as the key algorithm, which is related to the main topic of this lecture, and meets the requirement of Image Understanding and Recognition. Moreover, the result of this project can also be used to refine the embedding generated by the vision-based pipeline, and finally improve the downstream tasks.

Dataset

We use the Multitarget TED Talks Task (MTTT) dataset[3] as the original paper. The dataset is a collection of 30 TED Talks. It is now used as a *friendly competition* for machine translation. All the data are originally spoken and transcribed in English, then translated by TED translators. This dataset consists of 20 languages, i.e. 20 parallel pairs of translation. Each language contains a training set, a developing set, a testing set, and a set for metadata.
For this text recognition task, we use the English part of the *en-de* set[4] as the text-based data. Although this dataset is used for the machine translation task, the English texts are clean and abundant for our task. The statistical information of this set is as below:

Table 1: Statistical information of the English part in en-de set in the MTTT dataset

Set	Number of Sentences	Number of Tokens
Training set (/en-de/tok/ted_train_en-de.tok.clean.en)	151,627	3,037,569
Developing set (/en-de/tok/ted_dev_en-de.tok.en)	1,958	38,438
Testing Set (/en-de/tok/ted_test1_en-de.tok.en)	1,982	36,499

With this dataset, we will add noise into the text (especially the word-level granularity), and render them to an image as the input of the CNN-based neural network. The original text (token) before being polluted by the noise will be used as the golden label.

Approach

As described above, we will use a vision-based method for the text recognition task in order to improve the weak robustness compared to the text-based methods, and discuss the performance on robustness by selecting different font types.

[3] https://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/
[4] https://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/t/en-de/

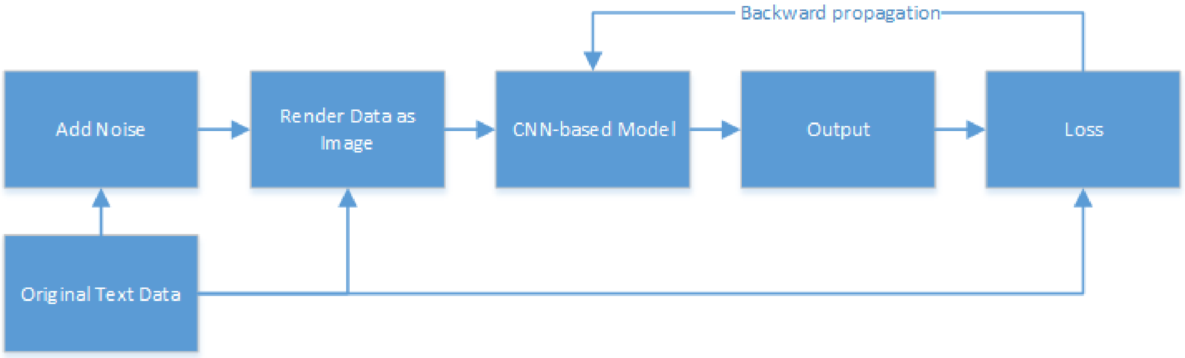

Figure 1

As is it showed in Figure 1, We now list the important steps:

1. extend the dataset by adding different percentages of noise like substitution,
2. build a pipeline to generate the input image for the CNN model. In this step, we need to select fonts other than Google Noto fonts to create comparison groups, render texts to images by *pygame* Python package,
3. establish an appropriate CNN-based model to process the rendered images and predict the text,
4. train the model, and tune the hyperparameters on both fonts to try to get the best results.
5. evaluate and compare with Precision, Recall and F1 score for both sets of fonts.

For the project-specific steps, we also need to:

6. write the final report,
7. record and remake the video.

Evaluation

The original paper showed the enhancement of the robustness of the text-related task by using the vision-based approaches. Based on this, we expect to see:

1. how good this approach handles the noised text recognition task with various proportion of noise,
2. whether the font plays a role to influence the vision-based robustness, - or how much a suitable font improves the robustness.

To validate our point of interest, we will compare the performance with different proportions of noise and different fonts, and evaluate their Precision, Recall and F1 scores, and then we analyze which type of fonts the model performs better based on these metrics.

Hardware

We are now able to access the following hardware (GPU) resources:

Table 2 Hardware Resources

Hardware	Number
NVIDIA GeForce GTX 1060 (12GB) personal laptop	1

NVIDIA GeForce GTX 1650 (12GB) personal laptop	1
Apple M1 Pro with 10-core CPU and 16-core GPU	1
NVIDIA T500 (12GB) personal laptop	1
NVIDIA GeForce RTX2080Ti (12GB)	4

Besides, we can also use resources from Google Colab and CIP.