11-777 Report 1: Dataset Proposal and Analysis

Hao Wu* **Jiayi Shen*** **Yanlin Feng*** **Yinghuan Zhang*** **Yuwei Wu***
{haowu3, jiayis2, yanlinf, yinghuan, yuweiwu}@andrew.cmu.edu

1 Problem Definition and Dataset Choice (1 page)

If you are choosing a dataset not listed on the course website, this section should be long enough to justify that you are qualified for your choice. This may mean a second page.

Dataset: TextVQA

1.1 What phenomena or task does this dataset help address?

This dataset helps address the task of visual question answering that requires reading and reasoning over the text in images, which is commonly needed by the visually impaired.

1.2 What about this task is fundamentally multimodal?

This task involves visual and text modalities. Solving this task requires the model to leverage both visual and text content.

1.3 Hypothesis

We believe there are three places cross-modal information can be used or improved

1. ...
2. ...
3. ...

1.4 Expertise

We have the following expertise in the underlying modalities required by this task:

1. Hao Wu: Took CV in Fall 2021, Experience in 3D vision
2. Jiayi Shen:
3. Yanlin Feng:
4. Yinghuan Zhang:
5. Yuwei Wu:

*Everyone Contributed Equally – Alphabetical order

2 Dataset Analysis (1 page)

2.1 Dataset properties

(GBs, framerate, physical hardware platform, ...)

TextVQA v0.5.1 contains 45,336 questions based on 28,408 images.

1. Training set contains 34,602 questions (103 MB) based on 21,953 images (6.6 GB) from OpenImages' training set.
2. Validation set contains 5,000 questions (16 MB) based on 3,166 images from OpenImages' training set.
3. Test set contains 5,734 questions (13 MB) based on 3,289 images (926 MB) from OpenImages' test set.

2.2 Compute Requirements

1. Files (can fit in RAM?)
2. Models (can fit on GCP/AWS GPUs?)

2.3 Modality analysis

(use a small sample – e.g. validation splits):

1. Lexical diversity, sentence length, ...
2. Average number of objects detected per image
3. Degrees of freedom, number of articulated objects, ...

2.4 Metrics used

2.5 Baselines

Four papers that have worked on this dataset

1. (Singh et al., 2019) is the original paper that introduced TextVQA dataset and the Look, Read, Reason & Answer (LoRRA) architecture
2. (Hu et al., 2020) has proposed a model for the TextVQA task based on a multimodal transformer architecture accompanied by a rich representation for text in images
3. (Kant et al., 2020) has proposed a novel spatially aware self-attention layer such that each visual entity only looks at neighboring entities defined by a spatial graph. Further, each head in the multi-head self-attention layer focuses on a different subset of relations

3 Team member contributions

Hao Wu contributed ...

Jiayi Shen contributed ...

Yanlin Feng contributed ...

Yinghuan Zhang contributed ...

Yuwei Wu contributed ...

References

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa.

Yash Kant, Dhruv Batra, Peter Anderson, Alex Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially aware multimodal transformers for textvqa.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.