Abstract

Since the advent of the printing press, text has slowly made inroads into the world we have built for us. The symbolic nature of text allows it to explain ideas more succinctly. Thus scene text content is often naturally occurring in images (street or storefront images). Further, they are also embedded into images to drive home clear takeaway points (e.g., printed posters, advertisement images). In both cases, though, they bring in crucial contextual information that aids in interpreting such images. However, despite this pervasion of scene text in our everyday images [Dey *et al.*, 2021] and the rich information source they entail, early works in visual understanding tasks like Image Classification, Captioning, and Visual Question Answering (VQA) [Antol *et al.*, 2015] did not leverage the scene text content of images. This can be attributed to the challenges of detecting and recognizing scene text in the wild. However, maturing research in Scene text recognition has improved their ability to read the text in natural images, thus making the scene text content more accessible. This easy accessibility of scene text content, coupled with the recent advances in multimodal architecturesHu *et al.* [2020], provides a unique opportunity to incorporate scene text into visual understanding tasks.

As our first point of the investigation [Dey *et al.*, 2021], we propose to jointly use scene text and visual channels for robust semantic interpretation of images. We not only extract and encode visual and scene text cues but also model their interplay to generate a contextual encoding with rich semantics. The contextual encoding thus generated is applied to retrieval and classification tasks on multimedia images with scene text content, to demonstrate its effectiveness. In the retrieval framework, we augment the contextual semantic representation with scene text cues to mitigate vocabulary misses that may have occurred during the semantic embedding. To deal with irrelevant or erroneous scene text recognition, we apply query-based attention to the text channel. We show that our multi-channel approach, involving contextual semantics and scene text, improves upon the absolute accuracy of the current state-of-the-art methods on Advertisement Images Dataset by 8.9% in the relevant statement retrieval task and by 5% in the topic classification task. Our results confirm our initial hypothesis that scene text plays an essential role in the semantic understanding of images. These results encourage us to extend our framework to more challenging tasks, like Text-VQA Singh *et al.* [2019a], that explicitly require us to read and reason with the scene text of an image. However, the scene text words come from a long-tailed distribution, giving such tasks *zero-shot* characteristics. We hypothesize that the zero-shot nature of these tasks can benefit from leveraging *external knowledge* corresponding to the scene text.

The open-ended question answering task of Text-VQA often requires reading and reasoning about *rarely seen or completely unseen* scene text content of an image. We address this zero-shot nature of the task by proposing the generalized use of external knowledge to augment our understanding of the scene text. We design a framework Dey *et al.* [2022] to extract, validate, and reason with knowledge using a standard multimodal transformer for vision language understanding tasks. Through empirical evidence and qualitative results, we demonstrate how external knowledge can highlight instance-only cues and thus help deal with training data bias, improve answer entity type correctness, and detect multi-word named entities. We generate results comparable to the state-of-the-art on three publicly available datasets under the constraints of similar upstream Optical Character Recognition (OCR) systems and training data. Through our experiments, we observe that this external knowledge not only provides invaluable information about unseen scene text elements but also augments the understanding of the text in general with detailed verbose descriptions. Our knowledge-enabled model is robust to novel text, predicts answers with improved entity type correctness, and can even recognize multi-word entities. However, the knowledge

pipeline is susceptible to erroneous OCR tokens, which can lead to false positives or complete misses. This also explains how our performance on the datasets is correlated with the particular OCR systems used.

Our investigation highlights the challenges and benefits of incorporating scene text into image understanding tasks. We validate our various hypotheses through empirical evidence across five different publicly available standard datasets. We conclude with a discussion on the implicit bias in these datasets for scene text, and propose data augmentation and a novel training scheme to deal with it.