1
Introduction

Images are the prevalent choice of expression these days, as they are often more engaging and less intrusive than other media. Often images use embedded text or scene text, in addition to visual elements, to express ideas more lucidly. Such images with scene text are ubiquitous in everyday life [Dey *et al.*, 2021], in the form of street or store-front images [Karaoglu *et al.*, 2017a] to posters, propaganda bills, advertisements [Hussain *et al.*, 2017] or simply brand logos [Wang *et al.*, 2020]. The scene text content in such images is often crucial in interpreting the image. More importantly, the scene text often has complementary cues, which, together with the visual cues, can lead to improved image semantics.

However, early works in visual understanding tasks like Visual Question Answering (VQA) [Antol *et al.*, 2015] and Captioning [Bai and An, 2018] did not leverage the scene text. Scene text recognition was still in its early stages, and manual annotation of text is resource-intensive. Text detection and recognition frameworks have matured in recent times, with datasets [Karatzas *et al.*, 2015b] and models which deal with real-life scenarios like complex backgrounds [Liao *et al.*, 2017; Jaderberg *et al.*, 2016], irregular font sizes, or arbitrarily oriented text [Liu *et al.*, 2018; Liao *et al.*, 2019; Xing *et al.*, 2019; Qiao *et al.*, 2019]. These recent advances in scene text recognition [Chen *et al.*, 2021b] have improved their ability to read the text in natural images, making the scene text content more accessible. The underlying scene text in images, which has been inaccessible until now in most image understanding tasks, can now be leveraged to interpret images. The work mentioned in this thesis particularly aims at

- Contextual understanding of scene text and visual objects
- Empowering scene text with external knowledge

Section 1.1 focuses on the research gaps that motivated us to explore the role of scene text as a contribution to research. In Section 1.2 we articulate our problem statement and give a brief overview of our work done. It also entails key insights while experimenting with various datasets and literature. In Section 1.3 we highlight our contributions. Finally, Section 1.5 presents the organization of chapters in the thesis.

1.1 MOTIVATION

The use of scene text in image understanding thus far has been scarce and initially limited to fine-grained classification tasks [Bai *et al.*, 2017; Karaoglu *et al.*, 2017b,a]. These works treat visual and text features as separate channels and do not model the semantic relationships. We hypothesize that rich contextual semantics can be generated by exploiting the relationship between co-occurring text and visual objects. This contextual encoding has also been explored concurrently by others in new language vision tasks and datasets, like Text Visual Question Answering (Text-VQA) [Singh *et al.*, 2019a; Biten *et al.*, 2019] and Text-Caption [Sidorov *et al.*, 2020], which explicitly requires

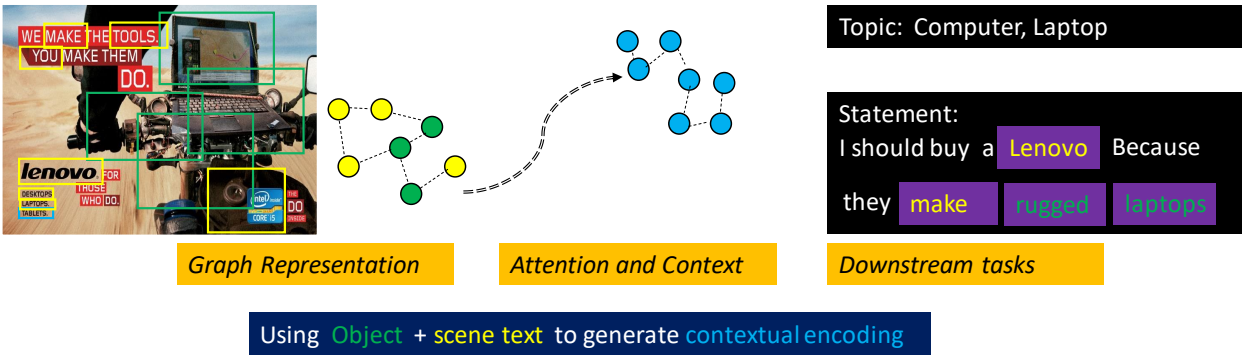

Figure 1.1 : Contextual Encoding– Apply attentional framework on a graph-structured organization of detected visual and scene text objects to generate contextual encoding. The said encoding is applied to downstream tasks like classification, retrieval, and even question answering.

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. In particular, we focus on the Text-VQA task, where understanding the scene text is necessary to answer questions about an image. The *zero-shot* nature makes it susceptible to errors due to *training data bias* when it fails to give due consideration to an instance only scene text cue and instead focuses more on trained statistics. E.g., in Figure 1.2, the standard Text-VQA model incorrectly predicts the answer to be 'samsung' because of the mobile phone-related visual content of the image. This observation motivates us to *empower* scene text, like 'vertu,' with *context appropriate* valid knowledge. Thus our second hypothesis is that the zero-shot nature of these tasks can benefit from leveraging *external knowledge* corresponding to the scene text. With the use of knowledge, we bring in contextual meaning about the unseen text 'vertu' as a mobile phone company which finally leads to the correct answer. We demonstrate more such cases in our experiments. While one can mitigate such dataset biases by additional training on large annotated datasets, we argue for the use of *external knowledge*.

1.2 PROBLEM STATEMENT

The goal of this thesis is the incorporation of scene text into image semantics. We propose two main ways to achieve this,

- We model the interplay between the detected text and the visual cues to generate a *contextual encoding* that can be applied to downstream tasks.
- We extend the contextual encoding with *external knowledge* to augment our understanding of scene text and deal with the long-tailed distribution of text. Unlike task-specific knowledge annotated datasets [Shah *et al.*, 2019b; Singh *et al.*, 2019b; Mishra *et al.*, 2019; Narasimhan and Schwing, 2018; Wang *et al.*, 2018], we leverage the availability of web scraped external knowledge base.

1.2.1 Overview of work done

This work proposes to go beyond detecting text and visual objects by learning a contextual semantic embedding that captures inter-object dynamics. Our contextual encoding scheme, as shown in Figure 1.1, incorporates visual objects with scene text to generate image semantic embedding. Our core idea is to represent an image as a structured graph of detected text and visual objects and then capture their interaction through attentional layers.

In our proposed Text-Visual graph, the detected text and visual objects are encoded alike using Convolutional Neural Network (CNN) generated appearance features and encoded word or class label features. This formulation allows features that are comparable and are in the same subspace. For our interaction scheme, we employ Graph Attention Layers (GAT) [Veličković *et al.*, 2017], which allows the nodes to generate task-specific contextual encoding. The encoded inter-object relationships, along with the feature encoding, augment the ability to reason about images.

The framework's ability to generate contextual encoding is evaluated across tasks and datasets. To show how the framework can adapt the contextual encoding to different scenarios, we apply the model to two datasets where context plays a critical role: advertisement images [Hussain *et al.*, 2017] and tweets [Gomez *et al.*, 2020]. We address two different tasks on the advertisement dataset (retrieval of relevant statements and topic classification) and a binary sentiment classification task (hate speech detection) on the tweet dataset. Both datasets contain images where text and visual elements are purposefully used to propagate an agenda, a marketing strategy, or a hateful message, as illustrated in Figure 2.1. They may also contain socio-cultural references, symbolism [Ye and Kovashka, 2018], along with wit and humor. Reasoning about such images involves understanding the context and the relationship between all the elements in that context [Zhang *et al.*, 2018].

For our classification task, the contextual encoding generated for the nodes is aggregated and then trained with FC(Fully Connected) classifier head. For the retrieval task, the aggregated node encoding is projected to semantic embedding space and trained with triplet loss with positive and negative statements. Further, we leverage the query statements' language structure by partitioning a statement into two parts and training two separate images to statement embedding. This enables us to model better the relation between the semantics of the query and the image. Experiments demonstrate that the use of scene text improves upon the state-of-the-art for both retrieval and classification tasks.

Our results show that the scene text channel is often more relevant than the visual channel for specific text-intensive datasets and tasks. Thus for our statement retrieval, we propose a multi-channel approach. In addition to the contextual semantic embedding, we also use text-only channels that provide semantic and lexical information. The improvements in results validate our hypothesis that scene text contains complementary cues, leveraging which leads to improved semantics.

While the scene text in contextual encoding leads to better semantic interpretation, it suffers due to the long-tailed distribution of the text. A rarely seen or unseen scene text word, which may be crucial to the current task, can often be overshadowed or completely ignored due to training bias. We propose using *external knowledge* corresponding to the scene text to address the zero-shot nature of these tasks.

The use of knowledge in Question Answering tasks has been proposed in the past but mainly applied to task-specific knowledge annotated datasets [Shah *et al.*, 2019b; Singh *et al.*, 2019b; Mishra

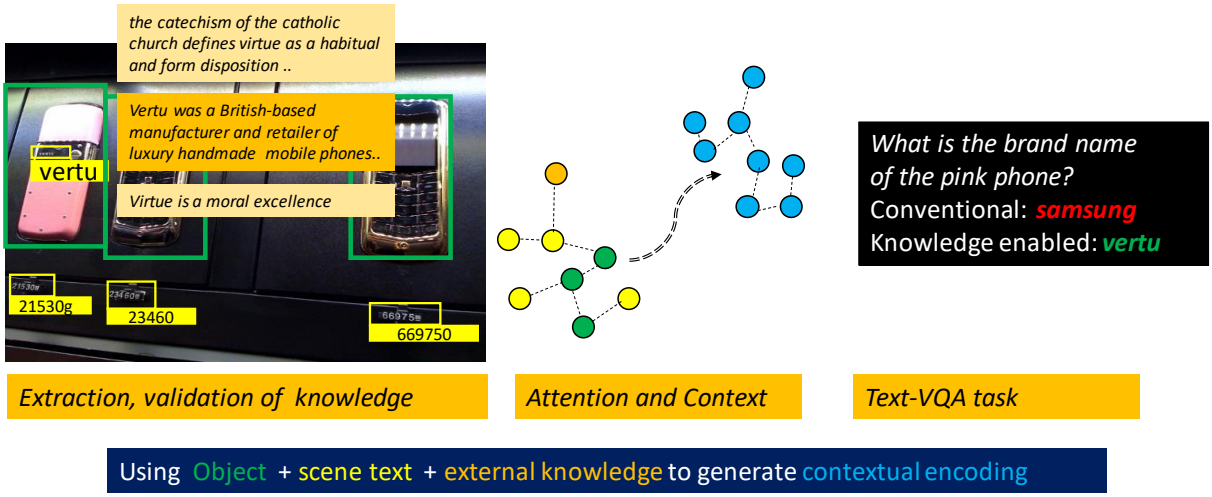

Figure 1.2 : Knowledge scheme – Use context validated external knowledge facts along with detected visual and scene text objects to answer questions about an image.

et al., 2019; Narasimhan and Schwing, 2018; Wang *et al.*, 2018]. We propose a more generalized way to add knowledge to any existing system. We leverage the availability of a web-scraped external knowledge base and apply it to existing Text-VQA datasets. Instead of relying on the availability of annotated knowledge facts, we extract multiple noisy candidate knowledge facts, which we validate through context before using. Furthermore, most works exploring the use of knowledge in VQA are set up to solve a *retrieval problem* [Shah *et al.*, 2019b; Narasimhan and Schwing, 2018; Wang *et al.*, 2018], where the answer is retrieved as the most relevant knowledge fact. The answers in Text-VQA are more diverse, extending beyond the retrieved knowledge facts. Thus we deal with noisy knowledge facts and produce answers by reasoning with them. We address the more general problem of what is valid and relevant knowledge and how to reason with it to generate the answer.

In Figure 1.2, we present our knowledge scheme applied to the Text Visual Question Answering (Text-VQA)task, which uses external knowledge along with visual and scene text cues to answer questions about an image. We propose an end-to-end knowledge processing pipeline that extracts and validates web scraped knowledge and reasons with it to generate the answer. We use the image context to filter out invalid candidate knowledge facts. Similar to the multimodal co-attention scheme of [Ye *et al.*, 2021], we propose a knowledge-enabled multimodal transformer [Vaswani *et al.*, 2017] to define task-specific relevance and reasoning with the validated knowledge facts.

Experimental results show that integrating knowledge in our framework leads to improved results with respect to the baseline in generic Text-VQA datasets, reducing errors due to data bias. However, Text-VQA is primarily a reading task and does not always require knowledge to answer questions. Thus, to validate the efficacy of our scheme in dealing with knowledge, we also apply our model to Text-KVQA [Singh *et al.*, 2019b], a knowledge-enabled dataset along the lines of Text-VQA, with *knowledge-oriented* questions *designed to require knowledge* to answer. Our results highlight how our approach of using mined external knowledge leads to better results than using the associated ground-truth knowledge facts included with the Text-KVQA dataset.

We conclude with a discussion on the bias for scene text. In particular, we examine the Text-VQA problem and highlight the inherent biases as part of the dataset design. We encountered such bias for scene text in our experiments, where it was often the better-recognized scene text

tokens that led to maximum improvements in performance. The bias for scene text in the design of question-answer pairs has also resulted in pre-training tasks that exploit them. Thus in chapter 5, we summarize our observations about bias and offer data augmentation and training schemes to counter them.

1.3 CONTRIBUTION

The use of scene text in language vision tasks is an active area of research at the time of writing this thesis. As such, there are overlaps in the research carried out by us with other research groups. In cases where our methods, developed independently and concurrently, could not match or improve upon the performance achieved by such other groups, we have chosen to adopt and refer to their works. We have taken utmost care in citing and referring to such concurrent works. An example of this would be our work on Text-VQA, where our GAT-based scheme could not match the performance of the transformer-based scheme introduced by [Hu *et al.*, 2020], and we have chosen to cite them and not include our work on Text-VQA as a standalone work. Our contributions are primarily in the incorporation of scene text into image semantics and in empowering the said scene text with external knowledge. Briefly, the different aspects of our contributions are as follows:

- We model the interplay between the detected text and the visual cues to generate a contextual encoding applied to different image understanding tasks, such as semantic retrieval, topic classification, and sentiment analysis. We believe we were the first to propose this integration of scene text into the semantics of the image.

- We propose a multi-channel approach for retrieval of statements by combining our contextual semantic embedding with additional text-only channels.

- We leverage the query statements' language structure to encode the relation between the semantics of the query and the image.

- To the best of our knowledge, we are the first to propose the generalized use of external knowledge to existing Text-VQA datasets, where we show that some questions can be answered correctly only by retrieving and reasoning with external knowledge. Our *External Knowledge* enabled Text-VQA (**EKTVQA**) the framework integrates the knowledge channel with a transformer based architecture for the Text-VQA task.

- We present a method to extract external knowledge facts corresponding to scene text words in images and then validate the mined knowledge facts through image context. This validation stage helps us deal with the noisy nature of web scraped data.

- The proposed end-to-end trainable architecture enables us to reason with external knowledge. We propose using masked attention maps to control the interaction between the image components and the knowledge facts through the transformer layers.

- We examine the bias in Text-VQA for scene text. Through experiments, we demonstrate this bias and propose a data augmentation and training scheme to deal with it.

1.4 PUBLICATIONS

Part of the work presented in this thesis has previously been presented as the following publications.

Journal:

- Dey, Arka Ujjal, Suman K. Ghosh, Ernest Valveny, and Gaurav Harit. "Beyond visual semantics: Exploring the role of scene text in image understanding." Pattern Recognition Letters 149 (2021): 164-171.

- Dey, Arka Ujjal, Ernest Valveny, and Gaurav Harit. "EKTVQA: Generalized Use of External Knowledge to Empower Scene Text in Text-VQA." IEEE Access 10 (2022): 72092-72106.

Conference/Workshop:

- Dey, Arka Ujjal, Suman K. Ghosh, and Ernest Valveny. "Don't only feel read: Using scene text to understand advertisements." CVPR Workshop (2018): Towards Automatic Understanding of Visual Advertisements (ADS).

Other publications during PhD which are not part of this thesis:

- Dey, Arka Ujjal, and Gaurav Harit. "Generating synthetic handwriting using n-gram letter glyphs." In Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1-8. 2016.

- Dey, Arka Ujjal, AH Abdul Hafez, and Gaurav Harit. "Greedy Gaussian Process Regression Applied to Object Categorization and Regression." In Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1-8. 2018.

1.5 ORGANIZATION

In this chapter, the importance of the problem of scene text in image semantics is introduced. A brief overview of the motivation behind incorporating scene text with visual cues for contextual encoding is also discussed. Further, we propose using external knowledge to deal with the long-tailed distribution of scene text. We briefly comment upon the challenges in both generating this contextual encoding and in empowering scene text with knowledge. Our clear contributions to addressing those challenges are highlighted to the reader.

In chapter 2, we review the existing work that motivates our ideas for contextual encoding and external knowledge to improve image semantics. The application of scene text in classification and retrieval tasks is discussed, elucidating the need for contextual awareness and attentional interaction. We discuss the challenges of using scene text in Text Visual Question Answering tasks, owing to its multimodal nature of text-visual objects, and the biases due to training data. In particular, we highlight the zero-shot nature of question-answering tasks, owing to the long-tailed distribution of text, and propose using external knowledge to deal with it.

In chapter 3, we present our contextual encoding scheme, which models the interplay between text and visual objects. We start with our visual and scene text encoding schemes and finally present our attentional interaction-based Contextual Visual Text encoder. Our proposed scheme is applied to classification and retrieval tasks across datasets, generating state-of-the-art

results. Through ablation studies, we highlight the effectiveness of our proposed model components, viz., contextual encoding, multi-channel based retrieval, and separately trained branches based on the linguistic structure of queries. However, this contextual semantic embedding framework often has to deal with rarely or completely unseen scene text tokens during test time.

In chapter 4, we propose using external knowledge to deal with long-tailed distribution of scene text. We present an end-to-end framework for extraction validation and reasoning with external knowledge. Instead of annotated knowledge facts, we argue for a knowledge extraction validation scheme that leverages freely available web-scale data for annotating our scene text tokens with multiple candidate knowledge facts, which are later filtered through image context. Our attention-based reasoning scheme is constrained to enforce knowledge legitimacy where knowledge facts can only influence the scene text query it originated from. Through experimental results, we showcase how our knowledge-enabled framework can reason with rarely or completely unseen scene text tokens.

In chapter 5, we study the implicit bias in Text-VQA for scene text. We identify scene text reference in the questions as a bias that lends it characteristics of a text localization problem against the intended reasoning problem it was supposed to be. We offer ways to mitigate such bias through data augmentation and novel training schemes. Further, we discuss the scene text dependence for answering. To elucidate this, we show through experiments how most questions can be answered only by looking at the scene text cues.

In chapter 6, we summarize the research work done in this thesis and provide a conclusion with valuable insights about the underlying challenges in incorporating scene text. Our contributions pertaining to visual-text contextual encoding and further extending it to use knowledge are highlighted. Finally, we conclude with possible future extensions involving scene text-based contextual OCR correction and external knowledge-guided explainability.