Conversational Question-Image Co-Attention for Visual Question Answering

Alvin Wan

Keywords: visual question answering, attention, computer vision

I. Introduction

Visual Question Answering (VQA) is a task that fuses both computer vision and natural language processing, where a computer must answer questions about a provided image. State-of-the-art approaches leverage attention networks in deep learning, which help to focus the neural network on portions of the image. In particular, these approaches see improved performance when applying attention to both the question and the image [1]. Further improvements are supported by use of external knowledge bases.

Related Work

Attention networks can currently leverage external information by captioning the image, extracting a set of attributes, and supplanting the final answer-generator with summaries of facts related to those attributes. However, what if the provided caption already ignores important context in the image? A caption may ignore a beach ball in the background, but a pointed question may not. What's more--what if additional context in the conversation is needed? The user may ask for "the number of people wearing jeans" and "*of those people*, the number of people wearing red shirts". I propose broadening the scope of information incorporated, when querying for external information, and applying attention to all the data retrieved, to build more stateful, conversational VQA bots.

II. Proposition

To broaden the scope of information incorporated, I propose running object detection on the original image to extract all identifiable entities; this can be achieved with a pretrained model, such as YOLO9000. The set of attributes obtained from captioning the image as well as describing these objects could then be fed into external information queries.

Provided that attention has improved performance when applied to questions and images alike, coupled with the fact that leveraging external information likewise improves performance, I propose applying attention to external information as well. Once performance has been verified, I then treat previous background in the sequence of questions, as external information. This then allows the VQA bot to utilize previous context.

In preparation for real-time inference and memory constraints, we will additionally consider size and space-efficient methods for accomplishing these improvements, so that the bot is able to build conversation.

Evaluation

A number of public benchmarks provide adequate testbeds for both more higher-level processing and more rote, simple questions. The COCO-QA and DAQUAR benchmarks would serve this purpose, additionally providing comparison with existing state-of-the-art methods. Other benchmarks that address subsets of the VQA task, such as image captioning, exist for intermediate evaluation.

[1] Lu et. al. "Hierarchical Question-Image Co-Attention for Visual Question Answering" (2016)

III. Research Plan

Year I: Question Attention, Broadened Information Retrieval

Objective: Test a number of varying implementations for raw information-based attention.
Methodology: Employ object detection and image tagging to increase the quantity of information queries. We can use both shallow and deep featurizations to produce a rich set of data descriptors. Then, apply attention to external sources to determine both quality and relevance of external information--we can begin by modeling relevance as a binary classification problem, then by regressing to continuous-valued relevance.

Year II: Stateful Visual Question Answering

Objective: Utilize previous conversation context to aid in question answering.
Methodology: Use attention networks to determine what types of external information are needed, whether the information pertains to objects, an intangible quality, or previous conversation context. Attention networks should successfully discern relevance of samples collected prior to learning. This work may extend to year three, as various modalities of external information may require significantly more data processing and filtering.

Year III: Towards more Conversational Bots

Objective: Run such evaluation and answering in real-time.
Methodology: This interest directly stems from my current work in reduced model sizes and real-time inference for self-driving cars. In a similar manner, we cannot always rely on network connections for high-level, fast-paced conversation. Neural network architecture and information retrieval both will be optimized by minimizing fully-connected layers, using convolutional layer reductions, and rigorous memory profiling. Attention networks should be trained and validated as part of an online learning algorithm.

IV. Conclusion

Intellectual Merit - Should this work succeed in proving an effective method of information distillation for incorporation, visual question answering as a field could take several steps forward in studying not just computer vision for images, or natural language processing for questions, but distributed systems for large-scale information retrieval and processing.

Broader Impacts - One application of an improved bot, with enhanced visual question answering capabilities, is personal assistants for the visually-impaired. Considering computer-aided descriptions are substitutes for the user's vision, the quality of visual question answering is highly-valued. However, the challenge is more complex than simply answering a number of isolated questions; the bot must be able to leverage the conversation context itself i.e., previous questions, the user's behavioral tendencies, and finally, external information that may make answers more accurate--for example, if a price tag is shown but the store-wide discount is announced online. Thus, with more refined methods of incorporating outside data, the bot could see major improvements in its value for the visually-impaired.