Deep Bayesian Network for Visual Question Generation

Badri N. Patro Vinod K. Kurmi Sandeep Kumar Vinay P. Namboodiri
Indian Institute of Technology, Kanpur
{badri,vinodkk,sandepkr,vinaypn}@iitk.ac.in

Abstract

Generating natural questions from an image is a semantic task that requires using vision and language modalities to learn multimodal representations. Images can have multiple visual and language cues such as places, captions, and tags. In this paper, we propose a principled deep Bayesian learning framework that combines these cues to produce natural questions. We observe that with the addition of more cues and by minimizing uncertainty in the among cues, the Bayesian network becomes more confident. We propose a Minimizing Uncertainty of Mixture of Cues (MUMC), that minimizes uncertainty present in a mixture of cues experts for generating probabilistic questions. This is a Bayesian framework and the results show a remarkable similarity to natural questions as validated by a human study. We observe that with the addition of more cues and by minimizing uncertainty among the cues, the Bayesian framework becomes more confident. Ablation studies of our model indicate that a subset of cues is inferior at this task and hence the principled fusion of cues is preferred. Further, we observe that the proposed approach substantially improves over state-of-the-art benchmarks on the quantitative metrics (BLEU-n, METEOR, ROUGE, and CIDEr). Here we provide project link for Deep Bayesian VQG https://delta-lab-iitk.github.io/BVQG/

1. Introduction

The interaction of humans and automated systems is an essential and increasingly active area of research. One such aspect is based on vision and language-based interaction. This area has seen many works related to visual question answering [1] and visual dialog [11]. Current dialog systems as evaluated in [9] show that when trained between bots, AI-AI dialog systems show improved performance, but that does not translate to actual improvement for Human-AI dialog. This is because, the questions generated by bots are not natural and therefore do not translate to improved human dialog. Therefore it is imperative that improvement in the quality of questions will enable dialog agents to perform well in human interactions. Further, in [20] the authors show that unanswered questions can be used for improving VQA, Image captioning and Object Classification. So the generation of natural questions will further improve performance on these tasks. While not as well studied as the other tasks of answering questions or carrying a conversation, there has been work aimed at generating natural and engaging questions from an image [38, 23] which is the VQG task. The underlying principle for all these methods is an encoder-decoder formulation. We argue that there are underlying cues that motivate a natural question about an image. It is essential to incorporate these cues while generating questions. For each image, there may be a different underlying cue that is most pertinent. For some images, the place may be important ('Is it a cowshed?') whereas for others the subject and verb may provide more context ('Are the horses running?'). Our work solves this problem by using a principled approach for multimodal fusion by using a *mixture of experts (MoE)* model to combine these cues. We hypothesize that the joint distribution posterior based on the cues correlates with natural semantic questions.

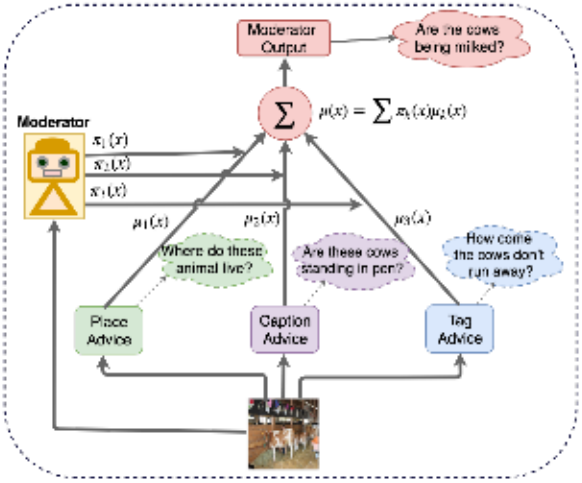

Figure 1. Here we give an overview of our network. We have three experts which provide us with information (advice) related to different cues. These are shown as Place Expert, Caption Expert and Tag Expert respectively. Then we have a moderator which weighs these advices and passes the resultant embedding to the decoder to generate natural question.

To verify our hypothesis, we systematically consider ap-