Some Questions to ask when designing Encoder-Decoder System for Image Captioning

Yogesh Tripathi, R. Pooja
CS16B044,CH15B056

May 12, 2019

Our project for the course, CS6370-Natural Language Processing, was implementation of Image Captioning System using Encoder-Decoder paradigm. However, we did not ask some interesting questions and mainly implemented an existing solution for the problem. So now we present some questions which were worth looking into while designing the system.

- **1. Looking closer into the encoder.** How can we make sure that the encoder is actually capturing only important concepts in the image? For example in Figure 1, identifying a market is more important than identifying a particular fruit/vegetable, which is more important than identifying the streetlight pole in the image.

Figure 1

  Recognizing what features are more important in an image, and at what level we should stop fine-graining requires World Knowledge based on how humans caption an image.

- **2. Looking into hardness in identifying concepts in an image.** Identifying different concepts in an image can be differently hard for an encoder. For example, it might be easier for the encoder to identify a concept like 'dance' as there might be features like costume, hands or legs moving, which distinguish dance from other concepts. However, identifying concepts like 'listening' can be far more difficult as one could be 'sitting', 'standing' or even 'dancing' and listening. There are very few distinguishing features for a concept like 'listening'. For such concepts, we can probably associate some relations, like identify listening if we identify speaking, or singing, etc and someone attending to it.

- **3. Looking into hardness in identifying different concepts associated with same word.** In some cases, the same word can have very different concepts in the image associated with it. For example, a concept like 'holding' can be very different in two images as in Figure 2.

Figure 2: Holding a bag vs Holding breath

Or we need not even go to such an extreme case, even holding a pen, holding a glass and holding a bag have very different visual representations. We should know the meaning of 'holding' something if we want to inject it into the captioning system or we should study what 'holding' exactly means to our learned model.

- **4. Language Grammar in Decoder.** As of now, our decoder has know clue about the grammar of the language. It generates grammatically incorrect sentences too as it is solely based on data. If we can inject some grammar information in the decoder, it would aid the decoder to generate more syntactically correct sentences.

- **5. Aiding Decoder by capturing relations between concepts using graphs.** We can use graph structures, mark concepts annotated in image by encoder as nodes and relations between them as edges. We can use this structure to help the decoder to build up a caption based on the graph structure.

This list is by no means exhaustive. However, looking into these aspects while developing the system can give insights into how humans see and describe, and how we can bring in the same kind of behavior in Captioning systems rather than going by simple data-driven approach, where mostly everything looks like magic in the end.