Towards Task Understanding in Visual Settings

Sebastin Santy, Wazeer Zulfikar
BITS Pilani KK Birla Goa Campus, India
f20150{357, 003}@goa.bits-pilani.ac.in
+1(412)708-5517, +91-7774031034

Rishabh Mehrotra
Spotify Research
London, United Kingdom
rishabhm@spotify.com

Emine Yilmaz
University College London
London, United Kingdom
emine.yilmaz@ucl.ac.uk

Abstract

We consider the problem of understanding real world tasks depicted in visual images. While most existing image captioning methods excel in producing natural language descriptions of visual scenes involving human tasks, there is often the need for an understanding of the exact task being undertaken rather than a literal description of the scene. We leverage insights from real world task understanding systems, and propose a framework composed of convolutional neural networks, and an external hierarchical task ontology to produce task descriptions from input images. Detailed experiments highlight the efficacy of the extracted descriptions, which could potentially find their way in many applications, including image alt text generation.

Introduction

A substantial portion of real world images depict a human task; for example, Figure 2 shows tasks like pitching a baseball, building lego toys or playing a guitar. While humans are efficient at understanding and describing the task intent by just a quick glance at a visual scene, most image captioning systems are only able to generate plain description of the different visual elements in the image. Especially in the cases of highly complex tasks, predicting task intent may involve more than just generating a plain description of the scene. Understanding the task context in visual scenes is indeed important in a number of application settings, including image alt text generation, image suggestions or image search.

While a lot of work has gone into generating image descriptions, most prior work (Bernardi and Cakici) have used visual and multi-modal space to assist the generation of a dense natural language description which allow for a more expressive prediction. Often times such detailed descriptions of the different visual elements are not required but a minimal explanation as to what is happening in the image suffices.Such *task*-based captions help in keeping the description more technical and contextual. For example, Pitch a Baseball is an apt minimal replacement to baseball player is throwing ball in game. in Figure 2, while being more technical, contextual and maintaining brevity. Our method primarily aims to improve existing methods by specifically overcoming this limitation by trying to precisely predict the task present in the scene while simultaneously preserving the context, as opposed to synthesizing a verbose description.

We jointly leverage insights from recent advancements in deep convolutional architectures and hierarchical task ontologies, and propose a two phase model to suggest scene task descriptions. The convolutional architecture generates contextual labels from the image, while the task extractor maps these labels to real world tasks. We leverage the TaskHierarchy138K [1] ontology which contains '*tasks*' and keywords associated with each of these '*tasks*' in a hierarchical structure, with a complex task often decomposed into simpler sub-tasks. Detailed experiments based on both qualitative and quantitative experiments demonstrate that our method not only helps in extracting task information, but also provides more useful descriptions when compared with state-of-the-art image description approaches.

Approach

In order to extract the tasks depicted in an image, we propose a two phased model: i) Multi-label classification of scenes to generate input labels for the task extractor and ii) Leveraging external hierarchical ontology for task identification by task extractor.

For image classification, we train the deep Inception v4 architecture (Szegedy and Ioffe) which is capable of detecting 1000 categories and fine-tune the network for multi-label contextual classification of the scene. The input image is fed to this classifier to obtain contextual labels along with their respective confidence scores. The labels generated by the classifier is further processed to filter out redundant information, and the resulting filtered set of labels is then passed to the task extractor module. Given the labels of the image, the task extractor will probe into a task hierarchy to suggest tasks . In order to infer the task, we leverage TaskHierarchy138K, which is an external task ontology that uses Wikihow articles to provide task information for over 200k real world tasks. Each node in this hierarchy represents a WikiHow category, with its children nodes representing its subcategories. A node k contains a Symptomatic embedding h_{Sk} and an Average Embedding h_{Ak}. h_{Sk} is an average embedding of the articles present in node k. h_{Ak} is
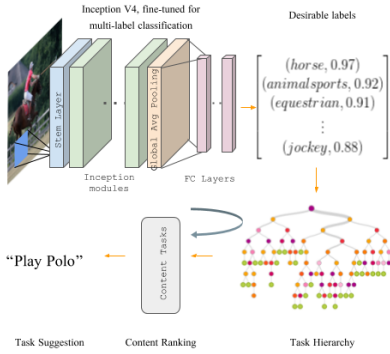
---

[1]https://usercontext.github.io/TaskHierarchy138K/

Figure 2: Description Comparisons

Figure 3: Crowdsourced evaluation compared

Figure 1: Architecture

$\frac{1}{n}\sum_{i=1}^{n=|children(k)|} h_{Ai}$ which is recursively calculated for all nodes except leaf-nodes. For leaf-nodes, $h_{Sk} = h_{Ak}$.

Given a task hierarchy and a set of labels produced from the classifier, we start with the root of the tree, and then trickle the labels down through the hierarchy. This trickling process is divided into two steps in order to achieve speed while simultaneously making it robust to noise.

1. **First Order Trickling**: We pass down a weighted average vector embedding (h_l) of all the labels starting from node k by recursively trickling to the child node with maximum cosine similarity ($cosine$) between h_l and $h_A \in children(k)$. The trickling stops at node k' when the $cosine(h_l, h_{Sk'}) \geq \forall cosine(h_l, h_A \in children(k'))$ where the $cosine(h_l, h_{Sk'})$ acts as a threshold.

2. **Second Order Trickling**: This is used for further trickling down the achieved node in the previous step. Some specific low-weighted labels belonging to a subcategory of the achieved node gets buried in h_l. Hence, we calculate $cosine$ between each label l and $\forall h_A \in children(k)$. The labels are trickled down to the node which returns the maximum $cosine$, iff it is higher than threshold defined in the previous step.

After the node is captured by the trickling process, we rank the content tasks using cosine similarity over the their respective article knowledge, to suggest the appropriate task.

Results and Discussion

To the best of our knowledge, this is the first work done on predicting tasks being undertaken in a given scene. Generating expressive image descriptions is the closest work to task suggestion. We compare our results with one of the best image descriptors in Figure 2. It should be noted that our work doesn't compete with the image descriptor. We want to make the reader aware that we are able to suggest task fairly accurately with a less complex model by leveraging existing task information.

For a qualitative analysis, we conducted two crowd-sourced studies on Amazon Mechanical Turk. For the first study, over 13 judges answer 10-randomly picked images along with image descriptions generated by NeuralTalk2(Karpathy and Fei-Fei), im2txt(Vinyals and Toshev
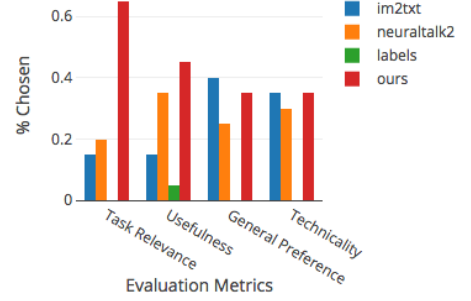
), multi-label classifier (Szegedy and Ioffe) (as baselines) and our method with each HIT being judged by atleast 7 unique workers. NeuralTalk2 uses convolutional and recurrent neural networks in multimodal space to generate image descriptions. im2txt is similar to NeuralTalk2 with a better classifier. We evaluate on the basis of 4 metrics: Task Relevance, Usefulness (with focus on brevity), General Preference and Technicality. As seen in Figure3, our method outweighs other captions for task relevance metric by a large margin (chosen more than other two combined) and performs almost equally for the other three. For the second study, we ask over 15 workers to rate our task suggestion on a scale of 1-5 independent of the baselines, giving appropriate measurement rubric to prevent biases. We achieve a score of 4.114/5. These two studies reinforce our assertion on task suggestion capability of our method.

Conclusion and Future Work

In this work, we propose a novel method for a scene task suggestion system. These descriptions can be used for applications like image alt text generation or as priors to existing image description models to build their descriptions upon, rather than generating them base up. However, this kind of a system is constrained to work on scenes where the task being done is a prominent part of it. We intend to extend this work to aid in the existing dense image description generation, making models intrinsically more task-aware by injecting task coherence scores within their architecture.

References

[Bernardi and Cakici] Bernardi, R., and Cakici, R. Automatic description generation from images: A survey of models, datasets, and evaluation measures.

[Karpathy and Fei-Fei] Karpathy, A., and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition.*

[Szegedy and Ioffe] Szegedy, C., and Ioffe, S. Inception-v4, inception-resnet and the impact of residual connections on learning.

[Vinyals and Toshev] Vinyals, O., and Toshev, A. Show and tell: A neural image caption generator.