**Motivation**

- Need for an understanding of the exact task being undertaken rather than a literal description of the scene.
- Leverage insights from real world task understanding systems, and propose a framework composed of convolutional neural networks, and an external hierarchical task ontology.
- Applications such as Image alt text generation.

**Approach**

In order to extract the tasks depicted in an image, we propose a two phased model:

1. Multi-label classification of scenes to generate input labels for the task extractor:
   - Inception Net used to produce labels.
   - Modified for multi-label classification.

2. Leveraging external hierarchical ontology for task identification by task extractor.
   - Task Hierarchy contains tasks/categories from Wikihow
   - Modified with insertion of word2vec embeddings at different levels to help with trickling of produced labels.
   - Two embeddings maintained at each node:
     - **Representative Embedding**: to describe the node characteristics in itself.
     - **Average Embedding**: to describe the children of a node. Calculated recursively on the representative embedding. Helps to avoid abstraction at higher levels of the hierarchy.
   - Trickling is based on the semantic similarity between the incoming labels and embeddings at each level.

**Results and Discussion**

- A crowd-sourced study on Amazon Mechanical Turk. In the study, workers answer 10-randomly picked images along with image descriptions generated by NeuralTalk2, multi-label classifier (as baselines) and our method.
- We evaluate on the basis of 4 metrics: Task Relevance, Usefulness, General Preference and Technicality.
- Our method outweighs NeuralTalk2 and im2txt captions for task relevance metric by a large margin.

**Conclusion and Ongoing 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.