OFA: UNIFYING ARCHITECTURES, TASKS, AND MODALITIES THROUGH A SIMPLE SEQUENCE-TO-SEQUENCE LEARNING FRAMEWORK

Nguyen Vu Duong[1,2,3]

[1] 20520465@gm.uit.edu.vn [2] University of Information Technology, Ho Chi Minh City, Vietnam [3] Vietnam National University, Ho Chi Minh City, Vietnam

Introduction

Achieving human-level competence across a universe of tasks and modalities remains a compelling but challenging goal in AI. To conquer this ambition, we propose a model that embodies three critical properties: task-agnostic, modality-agnostic, and comprehensive task coverage. However, existing single-modality and multimodal pretrained models often stumble upon these requirements. Therefore, we introduce a model, aspiring to become "One For All", unifying diverse input/output formats, tasks, and architectures.

Motivation

Recent advancements in the Transformer architecture and pre-training techniques show promise towards building an "omni-model" capable of handling diverse tasks and modalities. However, current approaches often violate key properties of an omni-model. They achieve this by introducing task-specific components, task-specific formulations, and entangled modality representations, which limit their ability to generalize and handle unseen tasks. This calls for a new approach that can overcome these limitations and build a true "one-for-all" model.

Overview

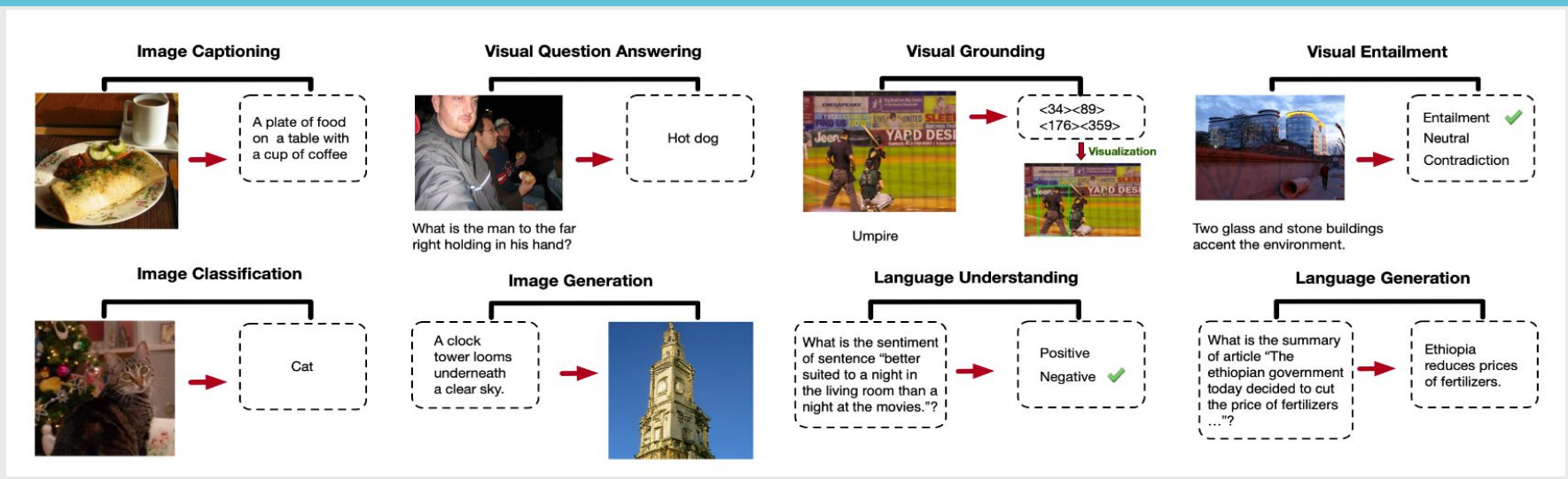

Figure 1. Examples of various tasks supported by OFA.

Description

1. Unified I/O

To enable flexible processing of various modalities without requiring dedicated output structures for each task, we leverage a unified vocabulary that encompasses all linguistic and visual tokens, including subwords, image codes, and location tokens.

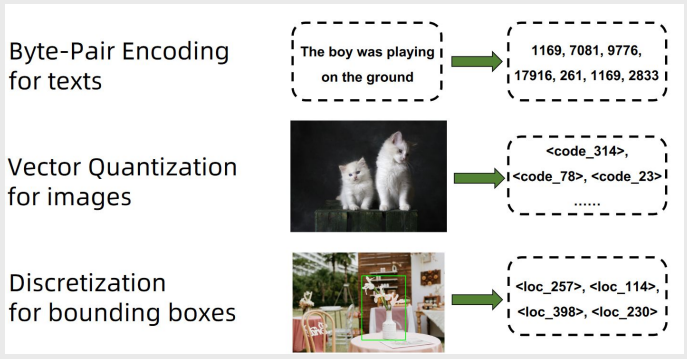

Figure 2. Modalities processing workflow.

2. Unified Architecture

- OFA utilizes a Transformer encoder-decoder architecture for processing various tasks. Notably, it avoids introducing any additional trainable components during both pre-training and fine-tuning phases.
- Furthermore, following Normformer, framework incorporates two additional Layer Normalization (LN) layers and Headscale attention mechanism to enhance training stability and accelerate convergence.

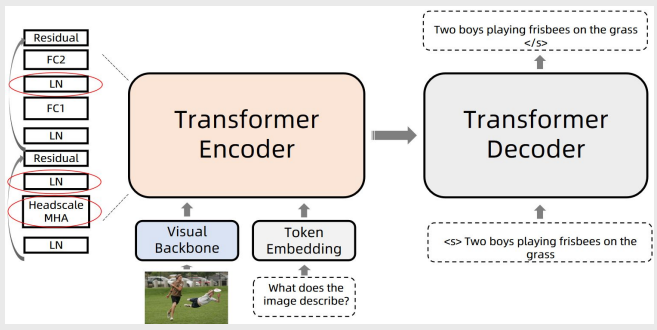

Figure 3. The demo architecture of OFA.

3. Unified Task

- Unifying tasks across modalities, we leverage a common sequence-to-sequence generation framework. To equip the model with comprehensive capabilities, we employ multi-task pre-training on multimodal and uni-modal data.
- We utilize handcrafted instructions to discriminate these tasks, further enabling model to perform zero-shot inference on new tasks.

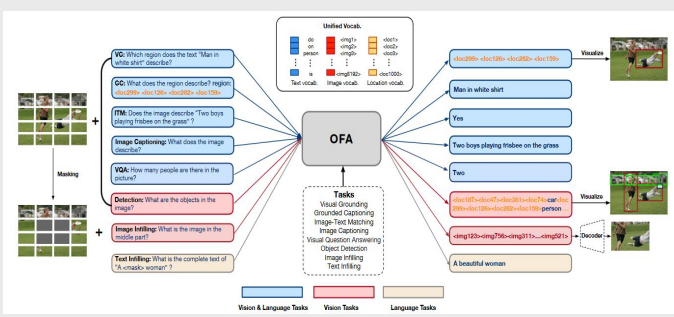

Figure 4. A demonstration of the pretraining tasks, including visual grounding, grounded captioning, etc.

4. Task Transfer

Research led us to develop a new task known as "grounded question answering," which requires the model to answer questions about specific regions of an image. This task proved to be an excellent test of the model's capabilities, and our results demonstrate its ability to achieve satisfactory performance.

Figure 5. Examples of Grounded Question Answering (unseen task)

5. Expected Results

- Build a complete model from the above ideas.
- The model solves multiple tasks but still achieves efficiency and even better than SOTA in each field.
- Report the methods and techniques of the developed model, experimental results, and evaluation
- A demo to perform a few tasks of the model to visualize.