Image Quality Assessment

Contents

Q-Align

Introduction

existing methods*(handcraft approaches, deep-neural-network-based methods)* :

- achieve remarkable accuracies on specific datasets by regressing from the **mean opinion scores (MOS)**.
- the complicated factors in contrast with the limited capacity of these methods have resulted in their poor **out-of-distribution (OOD)** generalization abilities. They usually experience compromised performance while handling different scoring scenarios together, making it challenging to train a unified model for different situations.

LLMs still fall short on accurately predicting scores that are consistent with human preferences:

- LMMs have similar behaviour patterns to humans while instructed to score: they prefer to respond with text defined levels *(good/poor)*; even while explicitly requested to predict numerical scores, the accuracy is significantly lower compared to deriving from levels.

Important one last mile: *How to teach LMMs to predict scores aligned with human?*

The standard process for collecting MOS from human:

- organizers need to define several rating levels (e.g. ‘excellent’, ‘fair’, ‘bad’) and select examples for each level, aligning human annotators to the standards of each level. human annotators never learns or marks a specific score. Instead, these final scores are derived from the distributions of human ratings.

Q-Align: a human-emulating syllabus to teach LMMs for visual scoring

- Training: simulating the process of training human annotators, we convert the MOS values to five text-defined rating levels*(excellent/*

good/fair/poor/bad), which are further formatted into instruction-response pairs to conduct visual instruction tuning on LLMs.

- Inference: simulating the strategy to collect MOS from human ratings, we extract the log probabilities on different rating levels, employ softmax pooling to obtain the close-set probabilities of each level. Get the LMM-predicted score from a weighted average on the close-set probabilities.

Q-Align

Methods

HOW DO HUMANS RATE?

1. **Training Human Raters**. Aligning human raters with one or more examples for each rating level. During this process, precise quality scores of the examples were not displayed to human raters.
2. **Collecting human ratings**. Collect initial human ratings. Human raters may provide their opinions in two types: 1) Directly choose rating levels. 2) Toggle the slider to generate a score.
3. **Converting human ratings to MOS**. Initial ratings are averaged into MOS in visual scoring datasets. Human raters do not participate in this stage.

HOW DO LMMS RATE?

instruction: `<img> Rate the quality of the image.`

Before specific alignment, LMMs predominantly respond with **qualitative adjectives**. Thus, if we use scores as the learning objective for LMMs, they need to first formally learn to output scores, and then learn how to score accurately. To avoid this additional formatting cost, we choose rating levels instead as the targets of Q-ALIGN.

Conversion between Rating Levels and Scores

[training] SCORES → RATIING Levels

Adjacent levels in human rating are inherently equidistant.

$$L(s) = l_i, \text{ if } m + \frac{i-1}{5} \times (M-m) < s \leq m + \frac{i}{5} \times (M-m)$$

$$\{l_i|_{i=1}^{5}\} = \{\text{bad}, \text{poor}, \text{fair}, \text{good}, \text{excellent}\}$$

[inference] RATING Levels → SCORES

reverse mapping: $G(l_i) = i$.

the MOS values are calculated via the weighted average of the converted scores and frequencies f_{l_i} for each level: $\text{MOS} = \sum_{i=1}^{5} f_{l_i} G(l_i)$.

For LMMs, we substitute the f_{l_i} with the LMM-predicted probabilities for each rating level.

Given that the predicted <LEVEL> token of LMMs is the probability distribution (denoted as χ) on all possible tokens of the language model, we conduct a close-set softmax[1] on $\{l_i|_{i=1}^{5}\}$ to get the probabilities p_{l_i} for each level, that p_{l_i} for all l_i sum as 1:

$$p_{l_i} = \frac{e^{\chi_{l_i}}}{\sum_{j=1}^{5} e^{\chi_{l_j}}}$$

the final predicted scores of LMMs:

$$S_{\text{LLM}} = \sum_{i=1}^{5} p_{l_i} G(l_i)$$

It represents the general expression form of the binary softmax

[1] I don't know why softmax is needed.

strategy ($S_{\text{Q-Bench}} = \frac{e^{\text{good}}}{e^{\text{good}}+e^{\text{poor}}}$) as proposed by Wu et al.(2023e)[2].

Model Structure

In the adopted structure, despite the visual encoder to convert images into embeddings, an additional visual abstractor further significantly reduces the token numbers per image (1024 $\rightarrow$ 64).

Under the 2048 context length for LLaMA2, we can feed as much as 30 images (2 without the abstractor) together during supervised fine-tuning (SFT). This allows us to input a video as a sequence of images to LMM, and unify image (IQA, IAA) and video (VQA) scoring tasks under one structure.

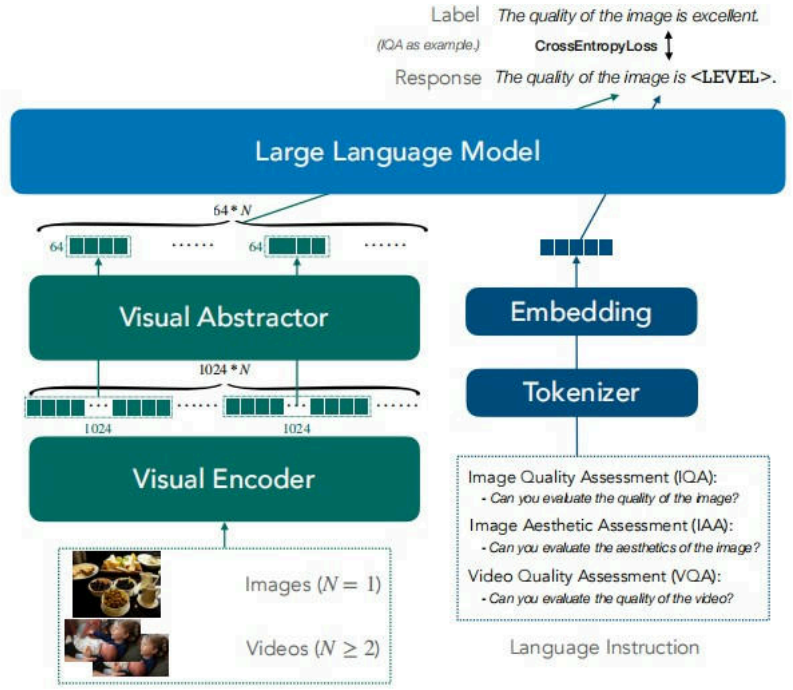

Figure 1: Model structure

[2]Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., and Lin, W. Q-bench: A benchmark for general-purpose foundation models on low-level vision. 2023e

Q-insight