QR-CLIP: Introducing Explicit Open-World Knowledge for Location and Time Reasoning

Weimin Shi [1] **Mingchen Zhuge** [† 2] **Zhong Zhou** [1] **Dehong Gao** [3] **Deng-Ping Fan** [4]

Abstract

Daily images may convey abstract meanings that require us to memorize and infer profound information from them. To encourage such human-like reasoning, in this work, we teach machines to predict where and when it was taken rather than performing basic tasks like traditional segmentation or classification. Inspired by Horn's QR theory (Horn, 1984), we designed a novel **QR-CLIP** model consisting of two components: 1) the **Quantity** module first retrospects more open-world knowledge as the candidate language inputs; 2) the **Relevance** module carefully estimates vision and language cues and infers the location and time. Experiments show our QR-CLIP's effectiveness, and it outperforms the previous SOTA on each task by an average of about 10% and 130% relative lift in terms of location and time reasoning. This study lays a technical foundation for location and time reasoning and suggests that effectively introducing open-world knowledge is one of the panaceas for the tasks.

1. Introduction

Many deep computer vision models have outstanding perception abilities and can solve regular tasks by extracting simple visual contexts (*i.e.*, color, texture, and objects), following the principle: "*what you see is what you get*". However, they cannot engage with a scene in the same insightful ways humans can (Crowder & Friess, 2012). It seems difficult for them to think deeper (or *learning to think*) based on their observations (Schmidhuber, 2015).

There are two reasons to explain the abovementioned problem: 1) until recently, many fundamental computer vision problems remained unsolved, so the community focused more on basic vision learning (Srivastava et al., 2015; He et al., 2016); and 2) previous models were unable to absorb more human knowledge due to limited hardware resources and data (Bommasani et al., 2021).

This paper aims to delve into the location and time reasoning behind the images (Fu et al., 2022). The procedure can be summarized as: *input an image; the goal is to have the model guess where and when the image was taken.* It is pretty different from the others (*e.g.,* basic classification (Wang et al., 2022), summarization (Fabbri et al., 2019), or retrieval tasks (Conforti et al., 2020)) since it requires the model explore more in-depth information and truly comprehend the event behind the images.

These days, industries afford millions of dollars to train foundation models (Reed et al.), and advanced parallel techniques (Rasley et al., 2020; Ott et al., 2019) enable the model to scale up the amounts of parameters and data. Some companies like OpenAI and DeepMind have developed a series of models, including GPT-3 (Brown et al., 2020), CLIP (Radford et al., 2021), and ChatGPT (Ouyang et al., 2022), *etc*. Most of those foundation models learn on their own from large amounts of online data made by people in a self-supervised fashion. This gives the models a certain amount of "open-world knowledge" (OWK). This motivates us to use CLIP (Radford et al., 2021) as our basic architecture to solve the proposed task since it shows effective performance in a number of multimodal tasks.

Compared with traditional image models (He et al., 2016; Liu et al., 2021), CLIP initially contains a certain amount of OWK. However, this knowledge is more encapsulated within the model. It can only be used implicitly through in-context learning (Min et al., 2022), preventing OWK from playing a larger role in our tasks. To solve this problem, we design a model named **QR-CLIP** with the capability to infer the location-and-time-related meta-information about the image. It is inspired by Horn's QR theory (Horn, 1984): Q-principle (quantity) requires maximization of the information content, which means "*ask speaker to present as much information as possible*". In contrast, the R-principle (relevance) requires minimization of form, which means "*should focus more on the relevant content*".

Equal Contributions: †Mingchen Zhuge designed and advised the project since March 2022. [1]State Key Laboratory of Virtual Reality Technology and System, BUAA [2]AI Initiative, KAUST [3]Northwestern Polytechnical University [4]CV Lab, ETH Zürich. Correspondence to: Zhong Zhou <zzhou@buaa.edu.cn>.