



_akhaliq

Following

Message



578 posts

334 followers

0 following

AK

POSTS

REELS

TAGGED

/visual Cortex and CortexBench

Website | Blog post | Paper



We're releasing CortexBench and our first Visual Cortex model, VC-1. CortexBench is a collection of 17 different AI tasks spanning locomotion, navigation, dexterous and mobile manipulation. We performed the largest and most comprehensive empirical study of pre-trained visual representations (PVFs) for Embodied AI (EAI), and find that none of the existing PVFs perform well across all tasks. Next, we trained VC-1 on a combination of over 4,000 ours of egocentric videos from 7 different sources and ImageNet, totaling over 5.6 million images. We show that when adapting VC-1 through task-specific losses or a small amount of in-domain data, VC-1 is competitive with the current state-of-the-art on all benchmark tasks.

Procedure-Aware Pretraining for Instructional Video Understanding

Honglu Zhou^{1,2}, Roberto Martín-Martín^{1,2}, Mubbasir Kapadia³, Silvio Savarese³ and Juan Carlos Niebles¹
¹Salesforce Research, ²Rutgers University, ³UT Austin
{hzhou, mml353}@cs.rutgers.edu, robertom@cs.utexas.edu, {ssavarese, jniebles}@salesforce.com

Abstract

Our goal is to learn a video representation that is useful for downstream procedure understanding tasks in instructional videos. Due to the small amount of available annotations, a key challenge in procedure understanding is to be able to extract from unlabeled videos the procedural knowledge such as the identity of the task (e.g., "make latte"), its steps (e.g., "pour milk"), or the potential next steps given partial progress in its execution. Our main insight is that instructional videos depict sequences of steps that repeat between instances of the same or different tasks, and that this structure can be well represented by a Procedural Knowledge Graph (PKG), where nodes are discrete steps and edges connect steps that occur sequentially in the instructional activities. This graph can then be used to generate pseudo labels to train a video representation that encodes the procedural knowledge in a more accessible form to generalize to multiple procedure understanding tasks. We build a PKG by combining information from a text-based procedural knowledge database and an unlabeled instructional video corpus and then use it to generate training pseudo labels with four novel pre-training objectives. We call this PKG-based pre-training procedure and the resulting model *PaperLika*. Procedure-Aware Pretraining for Instructional Knowledge Acquisition. We evaluate *PaperLika* on COIN and CrossTask for procedure understanding tasks such as task recognition, step recognition, and step forecasting. *PaperLika* yields a video representation that improves over the state of the art: up to 11.33% gains in accuracy in 12 evaluation settings. Implementation is available at <https://github.com/salesforce/paperlika>.

∞-Diff: Infinite Resolution Diffusion with Subsampled Mollified States

Sam Bond-Taylor, Chris G. Willcocks
Department of Computer Science
Durham University
{samuel.e.bond-taylor, christopher.g.willcocks}@durham.ac.uk

Abstract

We introduce ∞-Diff, a generative diffusion model which directly operates on infinite resolution data. By randomly sampling subsets of coordinates during training and learning to denoise the content at those coordinates, a continuous function is learned that allows sampling at arbitrary resolutions. In contrast to other recent infinite resolution generative models, our approach operates directly on the raw data, not requiring latent vector compression for context, using hypernetworks nor relying on discrete components. As such, our approach achieves significantly higher sample quality, as evidenced by lower FID scores, as well as being able to effectively scale to higher resolutions than the training data while retaining detail.

Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations

Jungo Kasai^{*}, Yubei Kasai[○], Keisuke Sakaguchi^{*}, Yutaro Yamada^{*}, Dragomir Radev^{*}
^{*}Paul G. Allen School of Computer Science & Engineering, University of Washington
[○]Sapporo Cardiovascular Clinic ^{*}Tohoku University [○]Yale University
jkasai@cs.washington.edu

Abstract

As large language models (LLMs) gain popularity among speakers of diverse languages, we believe that it is crucial to benchmark them to better understand model behaviors, failures, and limitations in languages beyond English. In this work, we evaluate LLM APIs (ChatGPT, GPT-3, and GPT-4) on the Japanese national medical licensing examinations from the past five years. Our team comprises native Japanese-speaking NLP researchers and a practicing cardiologist based in Japan. Our experiments show that GPT-4 outperforms ChatGPT and GPT-3 and passes all five years of the exams, highlighting LLMs' potential in a language that is typologically distant from English. However, our evaluation also exposes critical limitations of the current LLM APIs. First, LLMs sometimes select prohibited choices (禁選肢) that should be strictly avoided in medical practice in Japan, such as suggesting euthanasia. Further, our analysis shows that the API costs are generally higher and the maximum context size is smaller for Japanese because of the way non-Latin scripts are currently tokenized in the pipeline. We release our benchmark as JGAKU QA as well as all model outputs and examination data. We hope that our results and benchmark will spur progress on more diverse applications.

SoftCLIP: Softer Cross-modal Alignment Makes CLIP Stronger

Yuting Gao^{1*}, Jinfeng Liu^{1,2,*}, Zihan Xu^{1,*}, Tong Wu¹, Wei Liu², Jie Yang², Ke Li¹, Xing Sun¹
¹Tencent YouTu Lab ²Shanghai Jiaotong University
{yutinggao, lanxxu}@tencent.com, lijf19991226@sjtu.edu.cn

Abstract

In the preceding biennium, vision-language pre-training has achieved noteworthy success on several downstream tasks. Nevertheless, acquiring high-quality image-text pairs, where the pairs are entirely exclusive of each other, remains a challenging task, and noise exists in the data used datasets. To address this issue, we propose a novel approach that relaxes the strict one-to-one constraint and achieves a soft cross-modal alignment by using a softened target, which is generated from the self-supervised intra-modal self-similarity. The intra-modal self-similarity is indicative to enable two pairs have some local similarities and model many-to-many relationships between modalities. Besides, since the positive still dominates the softened target distribution, we disentangle the positive in the distribution to further boost the relation with the negatives in the cross-modal learning. Experiments demonstrate the effectiveness of SoftCLIP on ImageNet zero-shot classification and CC3M/CC12M as pre-training datasets. SoftCLIP achieves a top-1 accuracy improvement of 6.8%/7.2% CLIP baseline.

Language Models can Solve Computer Tasks

Guanyao Chen
University of California, Irvine
kguoc@uci.edu

Pierre Baldi
University of California, Irvine
pfbaldi@cs.uci.edu

Stephen McAleer^{*}
Carnegie Mellon University
smcaleer@cs.cmu.edu

Abstract

Agents capable of carrying out general tasks on a computer can improve efficiency and productivity by automating repetitive tasks and assisting in complex problem-solving. Ideally, such agents should be able to solve new computer tasks presented to them through natural language commands. However, previous approaches to this problem require large amounts of expert demonstrations and task-specific reward functions, both of which are impractical for new tasks. In this work, we show that a pre-trained large language model (LLM) agent can execute computer tasks guided by natural language using a simple prompting scheme where the agent Recursively Critiques and Improves its output (RCI). The RCI approach significantly outperforms existing LLM methods for automating computer tasks and surpasses supervised learning (SL) and reinforcement learning (RL) approaches on the MiniWoB++ benchmark. RCI is competitive with the state-of-the-art SL+RL method, using only a handful of demonstrations per task rather than tens of thousands, and without a task-specific reward function. Furthermore, we demonstrate RCI prompting's effectiveness in enhancing LLMs' reasoning abilities on a suite of natural language reasoning tasks, outperforming chain of thought (CoT) prompting. We find that RCI combined with CoT performs better than either separately.

Scaling Up Visual Speech Recognition With Synthetic Sup

Lakomkin², Konstantinos Vougioukas², Pingchuan Ma², Honglie Chen

SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer

Guanyao Chen^{1,2,*}, Zhijian Liu^{1,*}, Haotian Tang⁴, Li Yi^{1,3}, Hang Zhao^{1,3}, Song Han⁴

Ambiguity



Fine Details



MORE TO COME

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.