General Computer Interaction Language Alignment

Jacob F. Valdez
Limboid AI
jacob.valdez@limboid.ai

December 8, 2021

Abstract

This paper is written to fulfill the Term Project requirements for Data Mining.

1 Introduction

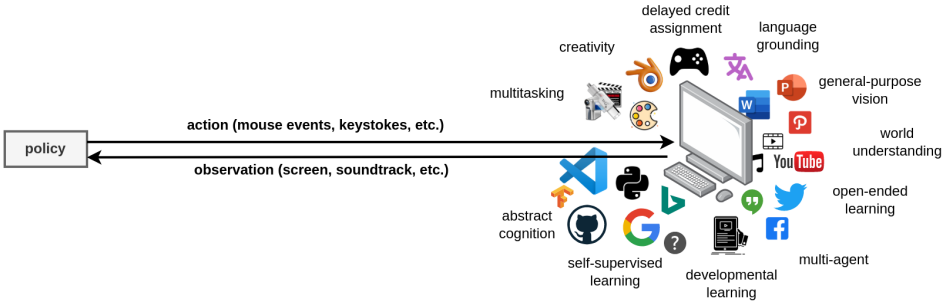

Figure 1: The general-purpose computer reasonably covers the anthropocentric problem domain. Performance across many tasks in this open-world domain therefore gives a proxy of development towards 'artificial general intelligence'.

The general-purpose computer provides a simple interface to vast distributions of natural and synthetic complexity which reasonably proxy the anthropocentric problem domain. This inherently includes any dataset machine learning practitioners might use, billions hours of recorded audio and video, live social media feeds, uncountable scientific, engineering, business, and historical documents, as well as creative software, integrated development environments, simulators, engineering design tools, e-commerce platforms, business systems, and many more applications. Considered together with the Internet, the general-purpose computer is a ready-made multiagent, language-grounded, lifelong-learning environment-incubator for the development-evolution of progressively more capable, general, and autonomous artificial intelligence.

Targeting this open set of tasks is not simple due to their non-stationary distribution. This is further complicated by heterogeneous user interfaces and context-sensitive application of natural world metaphors such as location, navigation, and gesture. Then there is also the issue of estimating task progress, completion, and reward in spite of shifting and overlapping task boundaries. While still keeping complete autonomy in mind as an ultimate objective, these challenges advocate occasionally relaxing the autonomy constraint in exchange for natural language human guidance.

Natural language is already ubiquitous across graphical user interfaces. It allows transferring not only objectives but also cognitive models from human to agent thus helping align both the agent's action and perception. Genuinely expressed natural language (not template statements) communicates deep relational hierarchies and dependencies. Most importantly, natural language is a high-bandwidth channel to rapidly infuse human-oracle information into the policy inference loop online. Rapid feedback accelerates the entire training loop iterating towards increasing capability, generality, and autonomy. Conversely, measuring a computer interaction agent's sustained alignment with natural language instructions over long trajectories may provide a reasonable proxy of development towards the illusion of artificial general intelligence. (See figure 1.)

This work represents one step in that direction. I introduce a heterogeneous multitask, multimodal semi-supervised dataset of recorded computer interactions – the

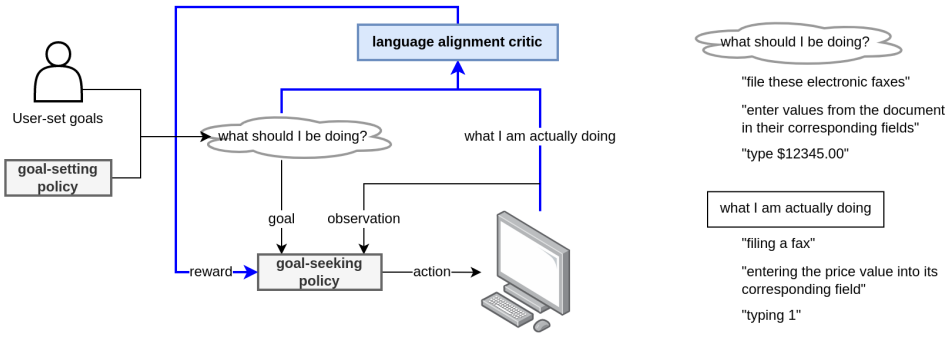

Figure 2: Overall architecture of a general-purpose computer interaction agent. (a) The language alignment critic (this work, colored blue) provides feedback to an online continually learning policy. (b) Goals are encoded are natural language statements at various levels of granularity such as “type $12345.00”, “enter values from the document in their corresponding fields”, and “file these electronic faxes”.

User Experience (UE) – and use it to train a multilevel action recognition system – the General Computer Interaction Language Alignment Critic (GCI-LAC). Future work will use this critic network to not only passively measure action-language alignment but also guide active inference (keystrokes and mouse actions) in a real computer environment. (See figure 2.)

2 The User Experience

Table 1: Composition of the User Experience

Dataset	Description	Modalities
COCO	image captioning	image, text
spoken mnist	spoken-written language	audio, text
synthesized ds	synthesized keystrokes and mouse events	keyboard, mouse, text

Computer interaction demands an understanding of diverse modalities: mouse events, keystrokes, language, audio, image, and video. At this scale of complexity, it is not currently feasible to build a massive supervised mouse-keyboard-text-audio-image-video dataset. Even if such a dataset were available, it may be unproductive to build training loops that demand every modality to be present in an example. For example, in many computer applications, the audio modality is ignored. It would be memory and compute efficient to similarly skip audio-related processing in corresponding dataset examples. However, in other applications such as media players, audio is essential and other modalities such as the keyboard and mouse can instead be ignored. Regardless of the modalities involved, this work aims to estimate a similarity measure between their current state and a natural language goal description. To my knowledge, no single dataset combines information from all these diverse modalities. Therefore, in this section I introduce a heterogeneous multimodal semi/supervised conglomerate dataset of datasets: the User Experience (UE).

The User Experience is currently composed of 3 datasets: COCO, spoken mnist, and a synthesized dataset of keystrokes and mouse events. Table 1 provides details on each of these classes. This collection will be expanded in the future. Some datasets in UE provide full descriptions at multiple levels of granularity, others pair brief or static inputs with single descriptions, and a large number merely provide raw data. Datasets are not batched by default. Each example is structured as a dictionary with the keys `mouse`, `keyboard`, `screen`, `audio`, `description`. Not all keys are present in every dataset example. The `mouse` modality is encoded by a 6-dimensional 32-bit float-valued tensor `<x location, y location, movement down (-) / up (+), movement left (-) / right (+), left button down, right button down>`. The `keyboard` modality is encoded as a 256-dimensional Boolean-valued vector with control, alphanumeric, and symbolic characters following ASCII mapping. The `screen` modality is variable sized RGB tensor with 32-bit floating point values already normalized in $[0, 1]$. The `audio` modality is encoded in a variable length 16kHz, 16bit normalized waveform with amplitude values in $[-1, 1]$. The `description` modality contains a concatenated string of natural language descriptions for the action, image, or audio that it is paired with. If there are no descriptions, this modality is dropped (i.e.: it always has a nonzero length). Sample rate varies between datasets. However all modalities except for `audio` share a common number of timesteps per individual dataset example. The audio modality, if present, will have significantly more entries on its time axis as a result of its 16kHz sampling rate.

3 General Computer Interaction Language Alignment Critic

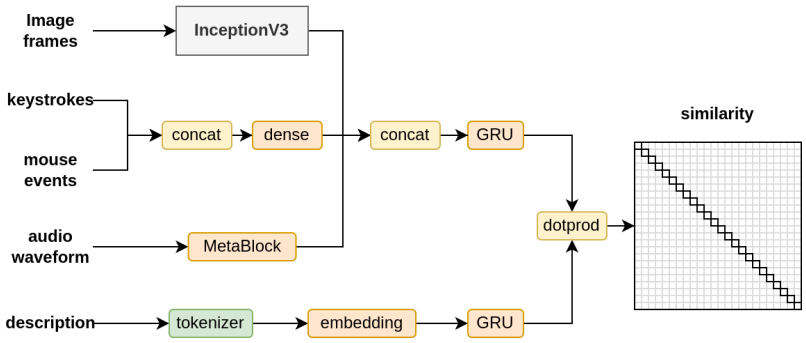

Figure 3: The language alignment critic uses a diverse set of modalities to predict a nontrivial vector that aligns with a language description semantic vector. Architecture primarily follows heuristic design.

As shown in figure 1, general-purpose computer interaction sits at the nexus of numerous problem domains involving mouse event and keystroke analysis, natural language processing, object detection, action sequence segmentation, audio/video understanding, and control. Drawing on existing contributions, this work combines pretrained models for most modalities separately and only trains a relatively small recurrent-state attention-based joint embedding network. The dot product between the joint embedding produced from computer modalities and the task semantic embedding is used to train a language alignment critic in CLIP-fashion. Figure 3 presents a visual anatomy of this architecture.