Implement Combination of Verbal and Non-verbal Communication for AI

Chenhao Zhou
Yuanpei College
Peking University
zhouch@stu.pku.edu.cn

Abstract

Non-verbal communication consists of eye-contact, facial expression, gestural communication and other forms of communication without language involvement. Several researches have shown that humans outperform other species in non-verbal communication for expressing complex intention. While Non-verbal communication, considered as a sign of human intelligence, shows ambiguous and less informative. In order to resolve the ambiguity and better implement the ability of non-verbal communication on AI, the essay discusses the alignment of verbal and non-verbal representations and long-term context information attention to improve the communication efficiency.

1 Introduction

Verbal and non-verbal communications are the daily communication patterns that humans are accustomed to. Liszkowski et al. [2] investigated that 12-month-old infants could gesture appropriately for knowledgeable versus ignorant partners, in order to provide them with needed information. It also shows that some basic yet effective forms of non-verbal communications such as pointing and waving contain abundant semantics and depend closely on context information. And without shared experience, joint attention, and common knowledge, the non-verbal representation itself hardly convey information.

In daily life, some verbal prompts could enhance the non-verbal communication efficiency and the process of conveying information is more robust. When single non-verbal gesture failed to express true intention, we will add some concise speeches to constrain the hypothesis space of corresponding intention. It inspires the combination of verbal and non-verbal communication patterns.

In this essay, we will focus on the gestures, thus discussing the alignment of verbal and non-verbal representations in Sec. 2. Then in Sec. 3 we try to promote a method that involving verbal prompts to assist the non-verbal communication. Finally we make a conclusion of the essay, which mainly focuses on the uncovered parts about communication.

2 Alignment both representations

Here we take open-domain dialog models as a specific example. The verbal and non-verbal representations can be extracted as the multi-modal representation. Thus the verbal and non-verbal communication is referred as a multi-modal scenario: the combination of linguistic and visual context. The rich multi-modal context can be involved during the alignment of verbal and non-verbal representations, which we will discuss based on two typical representation: joint and coordinated representation [1].

2.1 Joint representation

Joint representations combine the unimodal signals into the same representation space. Specifically for the non-verbal gesture and verbal linguistics context scenario, the same representation space can point to a hypothesis space of multiple intentions for the following considerations: 1) A non-verbal gesture would matches many verbal communications, towards many different intentions. For example, person A simply sticking out a finger at a cup and looking at person B, might convey the message like: "Fetch this cup for me", "This cup is for you" or even more requirement intentions. 2) The intention naturally contains a form of alignment between both representations. While with more long-term context involved in, the ambiguity can be resolved. For example, if we have additional information that is the shared experience, joint attention, or common knowledge such as: "The cup is empty", "Cups can hold liquids", it may be very likely that person A wants person B to fill the cup. Thus a soft alignment can be formed through the intention space based on the context.

Mathematically, the joint representation is expressed as:

$$x_m = f(x_1, \ldots, x_n, \mathbf{c}),$$

where the multimodal representation x_m is computed using function f that relies on unimodal representations $x_1, \ldots, x_n$ and complex context **c**, which refers to the measurement of shared experience, joint attention and common knowledge. While the joint representation is used to point to the intention or further action space.

2.2 Coordinated representation

An alternative to joint multi-modal representation is a coordinated representation. Instead of projecting the modalities together into a joint space, we learn separate representation for verbal and non-verbal expression but coordinate them through a constraint (obtained from the context). The alignment between gestures and speeches is formed through the context constraint. For the gestures as a unimodal information, some basic features which are the most relevant to representatinoal intent, can be extracted abstractly. For example, some of the gestures consisting of clear direction (indicating location), while others may contain thumb-up (indicating emotion). We hope these key features can be extracted while denoising and minimizing information loss. And the context information would be perceived for the alignment constraint.

Mathematically, the coordinated representation is expressed as:

$$f(x_1) \sim g(x_2),$$

where each modality has a corresponding projection function (f and g above) that maps it into a coordinated multi-modal space, but the resulting space is coordinated between them (indicated as $\sim$). The context information is involved in the coordinate process.

3 Combination of both representations

Srivastava and Salakhutdinov [3] identify additional desirable properties for multi-modal representations: it should be possible to fill-in missing modalities given the observed one. Thus some ambiguous non-verbal representations can be fixed to a certain extent by concise verbal prompts and vice versa. It is also consistent with human habits: we tend to combine the simple gestures with simple description to better convey our intentions.

In two major types of multi-modal representations we discussed above, joint representations are best suited for situations when all of the modalities are present during inference. But the coordinated representations project each modality into a separate but coordinated space, making them suitable for applications where only one modality is present at test time. Thus both of the representations can be benefited from the other modality prompt once the good alignment has been formed.

3.1 Conclusion

In short, the essay discusses how to implement the combination of verbal and non-verbal communication for AI from the perspective of multi-modality. However the rough problem abstraction would ignore some details, such as the appropriate representation of shared experience, attention and knowledge. The author still needs further exploration and study.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013. 1

[2] Ulf Liszkowski, Malinda Carpenter, and Michael Tomasello. Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, 108:732–739, 2008. 1

[3] N Srivastava and P. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *NIPS*, 2012. 2