Learning representation through communication

Olivier Tieleman, Angeliki Lazaridou, Shibl Mourad, Charles Blundell, Doina Precup
DeepMind

Human languages and their properties are greatly affected by the size of their linguistic communities. Small communities tend to develop more structurally complex languages, while larger communities give rise to simpler languages. [1, 2] At a more fine-grained level, speakers, aiming at maximizing communication effectiveness, adapt and shape their conceptualizations to account for the needs of their specific partners, a phenomenon termed in dialogue research as *partner specificity* [3]. In some extreme cases, the resulting *conceptual pacts* [4], are so ad-hoc and idiosyncratic that overhearers cannot follow the discussion [5].

In this work, we try to understand whether this community size effect on communicated representations is unique to humans and natural language, or also emerges in artificial learning systems. Our starting point is the traditional autoencoder (AE) which we view as a single encoder with a fixed decoder partner that must learn to communicate. We investigate whether *community-based learners*, which communicate with a multitude of partners (rather than a specific one), will shape the representations they communicate to be simpler in nature.

Community-based autoencoders (CbAE) consist of multiple encoders and decoders, which are paired up randomly at every training iteration to perform a traditional AE training step. Given that the identity of the decoder is not revealed to the encoder during the encoding of the input, the induced representation should be such that all decoders can use it to successfully reconstruct the input. A similar argument holds for the decoder, which at reconstruction time does not have access to the identity of the encoder. We conjecture that this process will reduce the level of idiosyncrasy, resulting in more abstract representations.

We apply CbAEs to two standard computer vision datasets and probe their representations along two axes, testing whether the community size effect results in learners that communicate *abstract* information of the images, such as *concepts* and their *properties*, rather than idiosyncratic and low-level visual information. We find that in contrast to representations induced within a traditional AE framework 1) the CbAE-induced representations encode concept-centric information that can be decoded by a linear classifier and 2) the underlying topology of the CbAE representations of concepts correlates better with human feature norms. Increasing community sizes turns out to reduce idiosyncrasies in the learned codes, resulting in representations that better encode concept categories and correlate with human feature norms.

[1] Matthew S. Dryer and Martin Haspelmath, editors. 2013. WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

[2] Gary Lupyan and Rick Dale. 2010. *Language structure is partly determined by social structure*. PloS one, 5(1):e8559.

[3] Susan E Brennan and Joy E Hanna. 2009. *Partner-specific adaptation in dialog*. Topics in Cognitive Science, 1(2):274291.

[4] Susan E Brennan and Herbert H Clark. 1996. *Conceptual pacts and lexical choice in conversation*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6):1482.

[5] Michael F Schober and Herbert H Clark. 1989. *Understanding by addressees and overhearers*. Cognitive psychology, 21(2):211232.