Grounding Words in Visual Perceptions: Experiments in Spoken Language Acquisition

Author: Fabio De Ponte[0000−0001−7139−3057]
Supervisor: Sarah Rauchas[0009−0008−2066−4222]

Department of Computing, Goldsmiths, University of London, Lewisham Way, New Cross, London SE14 6NW, UK

Thesis Abstract

In recent years, Natural Language Processing models have shown compelling progress in generating and translating text. Yet, the symbols that are manipulated by these models are not inherent to the models themselves. The machine only calculates the probability that a specific token comes after another (or a group of others) and then produces a list of tokens, each of which has a certain probability to follow the previous one. The semantic interpretation of the outputs – as well as of the inputs – of these models is completely invisible to the system that produces them. This is commonly referred to as the Symbol Grouding Problem and, as of today, there is not a generally accepted procedure to solve it.

Many attempts have been made at developing one. One of the earliest was proposed by Roy [1], who designed "a computational model which learns from untranscribed multisensory input" where acquired words were represented "in terms of associations between acoustic and visual sensory experience." The model was meant to mimic the way children learn, by discovering "words by searching for segments of speech which reliably predict the presence of visually co-occurring shapes." Another interesting experiment was proposed by Sugita and Tani [2]. They focused on the creation of a geometrical n-dimensional space, where the geometric arrangements represented "the underlying combinatoriality" among symbols. Yet another attempt came from Nolfi [3], who tried an evolutionary approach, by setting up an environment where communication emerged between behaviour-adapting robots after several generations. More recently, a number of other approaches were proposed. One came from Shao *et al.* [4], whose research was based on a learning from demonstration method through the "something something" video database [5] that offers images of thousands of actions on objects, paired with linguistic labels. A rather different approach came from Liu, Li and Cheng [6], who proposed a general-purpose neural sound synthesis (V2RA) network, based on generative adversarial networks (GANs), that was able to generate sound directly from visual inputs. In their work, partially recalling Roy's work, the task was "formulated as a regression problem to map a sequence of video frames to a sequence of raw audio waveform."

Based on this previous research, our project proposes another experiment, the training of a sequence-to-sequence model over videos characterised by visual elements which reliably predict the presence of acoustic co-occurring elements.

In order to perform the experiment, a dataset was created ad-hoc. It includes 5 types of objects (namely pen, phone, spoon, knife and fork) and 5 actions (move to the left, to the right, up, down and rotate). It is composed of 1,000 videos, each showing an object – either staying still or moving – while a voice reads a sentence, for example "this is a pen," that refers to what we see. There are twenty voices in total: ten voices are natural – i.e. they have been recorded by real people – while the other ten are artificial[1]. Each video is exactly 3 seconds long. It is 180x180 pixel sized and the audio is sampled to 16 KHZ. The data was augmented by five operations: flipping image, increasing and decreasing bright, increasing saturation and zooming in. In total, 36,000 samples were generated. The complete dataset is available online[2] and it is released under licence Creative Commons Attribution-ShareAlike 4.0 Generic (CC BY-SA 4.0). Video files were processed by two pre-trained neural networks, namely Wav2vec (a model introduced by Baevski *et al.* [7] at Facebook for self-supervised learning of representations from raw audio) and CLIP (developed by Radford *et al.* [8] at OpenAI), that extracted respectively acoustic and visual features. Then a sequence-to-sequence neural network mapped the extracted features of the visual part onto the extracted features of the acoustic part.

Two research questions were considered: whether such a model could map video features onto audio features, in fact producing a categorization without labels, where the categories would emerge from the parallel, simultaneous generalization of both input and target; and whether the model would be able to compose learned information about objects and movements to correctly describe a new combination, shown in a video it was not exposed to during training, a process that is defined as compositional semantics.

The experiment showed that the model was able to generalize simultaneously over videos and over the utterances that were paired with them. In fact, it produced sentences that were in some cases more accurate than the original ones, precisely because of the process of generalization. However, the results suggest also that the model did not develop the ability to combine information taken from different samples. In other words, while symbol grounding seems to have been achieved, compositional semantics does not.

The experiment shows that sensory perceptions can be mapped onto one another with a sequence-to-sequence model trained over a dataset where elements coming from different sensory domains are paired. However, it is not sufficient to develop compositional semantics.

Keywords: Computational Models of Human Learning · Symbol grounding · Compositional semantics · Natural Language Processing · Natural Language Understanding · Deep Learning · Fundamental research in AI · Human-centered AI · Robotics.

[1] Produced through the services of the website https://www.murf.ai
[2] https://www.kaggle.com/datasets/fabiodeponte/symbolgrounding

References

1. Roy, D.: Grounded speech communication. In: Proceedings of the 6th International Conference on Spoken Language (ICSLP), vol. 4, 69-72 (2000). Available at: https://www.isca-speech.org/archive_v0/archive_papers/icslp_2000/i00_4069.pdf. Last Accessed 22 May 2022.
2. Sugita, Y., Tani, J.: A sub-symbolic process underlying the usage-based acquisition of a compositional representation: Results of robotic learning experiments of goal-directed actions. In: 2008 7th IEEE International Conference on Development and Learning, ICDL (2008). Available at: https://ieeexplore.ieee.org/document/4640817. Last Accessed 16 September 2022.
3. Nolfi S.: "Emergence of communication and language in evolving robots" in Lefebvre, C., Comrie, B. and Cohen H., New Perspectives on the Origins of Language, 533–554 (2013). Available at: https://doi.org/10.1075/slcs.144.20nol. Last Accessed 8 May 2022.
4. Shao, L., Migimatsu, T., Zhang, Q., Yang, K., Bohg, J.: Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. In: The International Journal of Robotics Research, Vol. 40, Issue 12-14, October (2021). Available at: https://journals.sagepub.com/doi/abs/10.1177/02783649211046285. Last Accessed 16 September 2022.
5. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In: 2017 IEEE International Conference on Computer Vision (ICCV 2017) Venice, Italy. Available at: https://arxiv.org/abs/1706.04261. Last Accessed 14 September 2022.
6. Liu, S., Li, S., Cheng, H.: Towards an End-to-End Visual-to-Raw-Audio Generation with GAN. In: IEEE Transaction on Circuits and Systems for Video Technology, Vol. 32, Issue 3 (2021). Available at: https://ieeexplore.ieee.org/document/9430540. Last Accessed 6 September 2022.
7. Baevski, A., Zhou, H., Mohamed, A.R., Auli, M.: Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada (2020). Available at: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf. Last Accessed 14 September 2022.
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, A., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38 th International Conference on Machine Learning, PMLR 139 (2021). Available at: http://proceedings.mlr.press/v139/radford21a/radford21a.pdf. Last Accessed 14 September 2022.