Sam Acquaviva

Dec 2022

Establishing Compositionality in Infants

How do people form concepts? There are many hypotheses to answer this core question of cognitive science (Carey, 2011; Fodor, 1975; Levitin, 2002; Medin & Schaffer, 1978; Rosch, 1975). These theories differ wildly in describing which concepts are innate as well as how non-innate concepts are formed. However, they all share a common idea: concepts can be combined to form new concepts. For example, one can combine the concept of *green* with the concept of *couch* to form a new concept: *green couch*. This idea, that the meaning of an expression is a function of the meaning of its parts, is called *compositionality*.

Compositionality is incredibly important in understanding the world around us. For example, language is inherently compositional (Frege, 1991, p. 182-194). In a very similar vein to the example of the compositional concept *green couch*, the utterance "green couch" relies on the application of the adjective "green" to the noun "couch"; you must understand the meaning of the word "green", the meaning of the word "couch", and the relation between them to understand "green couch". To understand language, one can parse a sentence hierarchically into the meaning of its words. Similarly, to compose language, one must be able to compose the meaning of words into a sentence. These abilities, to decompose the speech of others and to use word meanings to generate more complex sentiments, inherently rely on compositionality (Matthei, 1982).

Even without language, people use compositionality. People can flexibly use the compositional structure of the visual world to understand scenes by parsing scenes into objects and the relations between them to make complex inferences or generate new visual concepts

(Zhou, 2021). In mathematics, compositionality is a well-defined concept that is a basis for describing more complex systems (Coecke, 2021).

Compositionality is also critically important in the field of artificial intelligence. For example, consider recent advances in large generative text-to-image models (Ramesh et al., 2022; Nichol et al., 2022; Saharia et al., 2022; Rombach et al., 2022). These models are large neural networks that are trained to generate an image from a text input. These models demonstrate the importance of scale, both in terms of dataset size and model size, on a variety of benchmarks. However, these models struggle to be consistently compositional. For example, see *Figure 1*, where the inferences from multiple state-of-the-art text-to-image models fail to compose the meaning of words to generate a coherent photo.

Figure 1: Randomly selected inferences from 3 state-of-the-art text-to-image models demonstrating lack of compositionality. *Left:* 16 images from GLIDE for the prompt, "a red cube on top of a blue cube". *Center:* 4 images from Imagen for the prompt, "A horse riding an astronaut". *Right:* 4 images from DALLE-2 for the prompt: "a red basketball with flowers on it, in front of a blue one with a similar pattern".

Other approaches attempt to improve the compositional ability of these by building in compositional priors into the architecture (Liu et al., 2022). Still, these methods are imperfect. Building in compositionality improves model performance and generalization when it is clear how to build in the hierarchy (Kuo et al., 2021), but in more complex domains (such as

text-to-image generation), we do not currently know how to correctly encode this hierarchy. The limits of these text-to-image models, despite increasing training data size and building in architectural priors, raise a broader question: can compositionality be learned?

Cognitive science does not have a clear answer to this debate. We cannot turn to humans to answer whether or not compositionality is innate. Currently, we do not know how early compositional reasoning emerges. Research has shown that infants can effectively use the principle of compositionality when dealing with language (Pomiechowska et al., 2019; Fernald et al., 2010). Pomiechowska (2019) showed that 12-month-olds can compose a known noun with a novel quantity label to compute the meaning of a complex noun phrase. For example, if it is physically demonstrated that "dax ducks" means "three ducks", infants can generalize that "dax balls" means "three balls". However, this ability does not necessarily generalize outside the language domain. Instead of exploring *linguistic compositionality* (composing the meaning of words to understand sentences), we are interested in *semantic compositionality* (composing the meaning of concepts to understand more complex concepts).

Piantadosi (2016) investigated whether 3.5-4.5 year-olds can apply compositionality in a visual setting. In this experiment, a function took the form of a screen that temporarily occluded a patterned object and changed a property of that object's pattern. Children were trained to predict the outcome of two distinct visual functions (two different screens that change different properties), $f(x)$ and $g(x)$. Then, at test time, they were shown an object going behind the two screens, and they were tasked with choosing the new pattern of the object ($f(g(x))$). The study found that they could predict the outcome of visual compositional functions at greater than chance, indicating that they can apply compositional reasoning.

However, Piantadosi (2018) found no evidence that 9-month-olds can understand compositionality in a similar experiment. In this study, however, they used a violation of expectation paradigm instead of a forced-choice paradigm. In contrast to the experiment with older children, the infants' looking times reflected that they mistakenly represented $f(g(x)) = f(x)$.

However, this failure to show compositional reasoning could be due to their limited working memory, rather than a lack of compositional understanding. Káldy (2005) suggests that the working memory capacity of 6.5-month-olds is limited to one occluded object if there is another object that is also occluded. Additionally, Feigenson (2009) suggests that 11-month-olds could not dynamically update a representation that was no longer in the current focus of attention. In tandem, these studies suggest that asking the 9-month-olds to apply two functions to an occluded object puts too much strain on their memory capacities.

As such, we aim to fill this gap in the understanding of compositionality in humans by probing compositionality in infants without requiring them to update internal representations beyond their limits. Specifically, instead of probing whether infants can understand $f(g(x))$, we instead ask can infants understand $f(x, y)$, where x is an object, f is a transformation of the object, and y is a visual parameter that changes the transformation f. In this setting, all components of the compositional function are visible during the entirety of the testing trial. We could present a patterned shape as x, a different patterned shape as y, and f as a function which applies the pattern of y to the shape of x. We can then test if infants reliably look longer at the output of $f(x, y)$ when it does not successfully apply the function, for different combinations of y and x. The experimental setup is shown in *Figure 2*. This is in contrast to the design of

Piantadosi (2018), where the functions f and g were presented in sequence, and the infant was tasked with updating the representation of x twice.

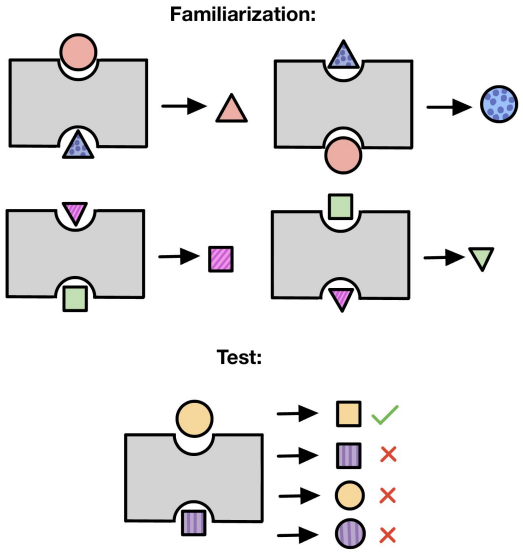

Figure 2: Familizariation and test trials for the condition where f applies the pattern of the top object x to the shape of the bottom object y. During the familiarization trials, the function f is demonstrated for both orderings of both pairs of objects. At test time, the baby is presented with 1 of 4 possible outcomes. If the infant can successfully learn the two-argument f, we expect the infant to look longer at the incorrect outcomes.

Methods

Subjects

Participants were 60 healthy infants, 34 male and 26 female, from the Greater Boston Area (mean=9 months and 12 days, range = 9 months and 2 days to 10 months and 12 days). All of the participants were from two-parent households, and most were from Caucasian, middle- to upper-class families. An additional 7 infants were tested but excluded from analysis for being inattentive (6) or displaying looking times more than three standard deviations from the mean (1).

Apparatus

The experimental apparatus consisted of a single screen (width=36", height=24") placed 3 feet in front of the infant. The screen showed animations involving the gray box and one of the three pairs of differently colored/textured shapes shown in *Figure 2*. During each of the trials, the gray box represents the function f, and the pairs of shapes shown below represent x, and y, the arguments of f.

Trials

The trials are depicted in *Figure 2*. There are 2 phases: the familiarization trials, and the testing trial. Whenever a trial introduced a new pair of objects, there would be a pre-trial to familiarize the infant with the objects. Each pre-trial consisted of an 8-second animation where each object is highlighted in turn as an audio recording of an experimenter played saying "look!".

Each main trial consisted of a 25-second animation. During the animation, the box depicting f always stayed at the center of the screen. Then, the object representing argument x would slide into the screen, where it would be highlighted as an audio recording of an experimenter played saying "look!". Then, x would slide into its slot in f and a cheerful "ding" sound would play. This sequence lasts 10 seconds. This was repeated for the object representing argument y. Then, a new object representing $f(x, y)$ would slide out to the right of f that has the shape of object y and the pattern of object x. This sliding animation lasted 5 seconds.

During the trials, there were two conditions. In one condition $f(x, y)$ would apply the pattern of x (the top object) to the shape of y (the bottom object). In the other condition, $f(x, y)$ would apply the pattern of y to the shape of x.

Familiarization

During the familiarization trials, the function f is demonstrated for various combinations of objects. Specifically, there were 4 familiarization trials with two pairs of objects. For both

pairs of objects, the function f would be demonstrated with both possible orderings of the pair. Before f was demonstrated for a pair of objects, there is a pre-trial where the infant is shown the pair of objects on the screen. An 8-second animation is played where each object is highlighted as an audio recording of an experimenter played saying "look!".

Test

During the test trial, the infants are shown a novel pair of objects as x and y, the arguments of f. Then, the 25-second animation is shown, and the looking times of the infants are recorded. There are four possible outcomes: 1). the correct application of the function ($f(x, y)$), 2). the application of the function in the wrong order ($f(y, x)$), 3). object x, and 4). object y. Each of these outcomes is shown in *Figure 2.*

The ordering of x and y are counterbalanced in different experiments so as to account for possible *a priori* preferences for different shapes.

Procedure

The infant was seated on a parent's lap 3 feet in front of the screen. The parents were instructed to keep their eyes closed and remain silent and neutral throughout the experiment.

Each session was recorded. The recording of each trial was then reviewed by three observers who indicated when the infant was and was not looking at the result on the screen. To assess interobserver agreement during the main trial portions of the familiarization, display, and test trials, each trial was divided into 100 millisecond intervals. For each interval, we computed whether all three observers agreed on whether the infant was looking at the event during each interval. The percent agreement for each trial was calculated by dividing the number of intervals in which the observers agreed by the total number of intervals in the trial. The average agreement

across all three raters for all participants was 95% per trial per infant. For our results, we used the average looking time recorded by these three labelers.

The infants initially underwent four familiarization trials with two pre-trials (before a new pair of objects is introduced). An examination of the infants' looking times during the pre-trial portion of each trial revealed that they were highly attentive to the objects, with an average looking time of 7.2 / 8 seconds. During the main trial portion of each familiarization trial, the trial ended when the infant looked away for two consecutive seconds after having looked at the objects for at least two cumulative seconds.

Finally, the infants were presented with one of four possible test trials, with each trial displaying an event appropriate for their assigned condition. The infants were highly attentive during the 8-second pre-trial at the start of each trial, with an average looking time of 6.9 / 8 seconds. The main trial portion of each test trial ended when the infant looked away for two consecutive seconds after having looked at the objects for at least two cumulative seconds..

Results

Analyses of the familiarization and test data did not reveal any significant effect of event condition, sex, or choice of test-trial object pair on looking times. This means that 1). infants tended to look equally long during trials where f applies the pattern of x to the shape of y versus trials where f applies the pattern of y to the shape of x, 2). male infants tended to look equally as long as female infants, and 3). the choice of which 2 pairs of objects to show during familiarization did not affect looking times. As a result, the data were combined across sex, event condition, and choice of test-trial object pair in subsequent analyses.

The infants' looking times during the main trial portions of the two test trials were averaged and analyzed. The infants tended to look longer at each of the two incorrect outputs

that generates one of the inputs ($f(x, y) = x$, $f(x, y) = y$) than the correct application of the function ($f(x, y)$) ($p < 0.01$) and the application of the function in the wrong order ($f(y, x)$) ($p < 0.05$). Infants also tended to look longer at the application of the function in the wrong order ($f(y, x)$) ($p < 0.05$).

Discussion

These results indicate that infants as young as 9 months old can accurately predict the output of a visual function with two arguments. This ability is evidence that infants can reason compositionally outside of language; the meaning of the function f is determined by its parts (x and y) and their combination (applying the texture of y to the shape of x as in condition 1, or applying the texture of x to the shape of y as in condition 2).

The positive result opens up a few new possibilities about the development of compositionality. Perhaps the ability to reason compositionally develops in domains outside of language *before* it develops in language. Prior to this work, the current literature would suggest the opposite. This proposal seems logical; if infants do have some sort of internal compositional representation, then it seems likely that they would be able to apply it to domains they have a firmer grasp on (such as shapes and colors) before applying it to domains they are still in the early stages of learning (such as language).

Another possibility is that infants can apply compositional reasoning from birth. Developing experiments to test this theory will be incredibly difficult because the memory and attentional limits of newborns are even more limited than 9-month-olds. Demonstrating that newborns can apply compositional reasoning would offer an interesting perspective to machine learning. Currently, state-of-the-art models in language modeling and in computer vision do not have any architectural biases toward compositionality. If humans have compositionality from

birth, then perhaps there should be more work into building in compositionality priors into these models.

Limitations

One major limitation is that this study presents just one single task. It is not evidence that infants can successfully apply compositional reasoning generally. It could be that visual transformations are like language in that infants demonstrate compositional reasoning in this domain before demonstrating it in others. Additionally, this compositional task is fairly simple. Whether or not infants can extend this ability to visual transformation tasks in the natural world is unclear. So, although our results show that infants can apply compositional reasoning in a very simple environment, their other cognitive constraints may limit its applicability; young infants may not use their compositional ability to understand the world around them.

Another limitation is that understanding the visual function $f(x, a)$ is not synonymous with compositionality. Although the alternative explanation to the null result of the 2018 Piantadosi study (that the null result stemmed from limited attentional resources rather than an inability to reason compositionally) is logical, we still do not show that infants can predict the outcome of $f(g(x))$. It is possible that infants can predict $f(x, y)$ without the ability to predict $f(g(x))$. Our experiment shows some level of compositional reasoning ability, but we cannot conclude that infants are completely compositional from this data, even in simple tasks.

More generally, we cannot conclude from experiments about compositional reasoning that infants have a compositional representation. There are cases in which behavioral data can provably show that a certain representation must exist (Piantadosi 2022), but proving that there is a compositional representation based on successful prediction of the outcomes of compositional challenges is much more difficult and not addressed here.

Conclusions

These results suggest that infants as young as 9 months old are able to accurately predict the outcome of a visual function with two arguments. This ability provides some evidence that infants can reason compositionally without relying on language; the meaning of the function f is derived from its parts (x and y) and their combination (applying the texture of y to the shape of x or vice versa).

This extends work by Piantadosi 2016, which shows that 4-year-olds can reason compositionally about visual functions, and Pomiechowska 2019, which shows that 12-month-old infants can reason compositionally on linguistic tasks. It provides some evidence that the null result of Piantadosi 2018 was due to attentional limitations of infants rather than an inability to reason compositionally. Future research should explore if younger children can correctly predict the outcome of 2-argument visual functions. Additionally, future work should explore other types of tasks that require compositional ability to see how general this reasoning ability is. Combining these two suggestions, if future work showed that performance on compositional tasks across domains (e.g. language, visual) emerged at the same age, then this would provide some evidence that there is some common compositional representation that is emerging.

References

1. Carey, Susan, (2011). The origin of concepts. Oxford University Press.
2. Coecke, B. (2021). Compositionality as we see it, everywhere around us. arXiv. https://doi.org/10.48550/arXiv.2110.05327
3. Feigenson, L., & Yamaguchi, M. (2009). Limits on infants' ability to dynamically update object representations. Infancy, 14(2), 244-262.
4. Fernald, A., Thorpe, K., & Marchman, V. A. (2010). Blue car, red car: Developing efficiency in online interpretation of adjective–noun phrases. Cognitive Psychology, 60(3), 190–217. https://doi.org/10.1016/j.cogpsych.2009.12.002
5. Fodor, Jerry A., (1975). The language of thought. New York: Thomas Y. Crowell.
6. Frege, G. (1991). Collected papers on mathematics, logic, and philosophy.
7. Káldy, Z., & Leslie, A. M. (2005). A memory span of one? Object identification in 6.5-month-old infants. Cognition, 97(2), 153–177. https://doi.org/10.1016/j.cognition.2004.09.009
8. Kuo, Y.-L., Katz, B., & Barbu, A. (2021). Compositional networks enable systematic generalization for grounded language understanding. arXiv. https://doi.org/10.48550/arXiv.2008.02742
9. Levitin, D. J. (Επιμ.). (2002). Foundations of Cognitive Psychology: Core Readings. doi:10.7551/mitpress/3080.001.0001
10. Liu, N., Li, S., Du, Y., Torralba, A., & Tenenbaum, J. B. (2022). Compositional visual generation with composable diffusion models. arXiv. https://doi.org/10.48550/arXiv.2206.01714

11. Matthei, E. H. (1982). The acquisition of prenominal modifier sequences. Cognition, 11(3), 301–332. doi:10.1016/0010-0277(82)90018-X
12. Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207
13. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv. https://doi.org/10.48550/arXiv.2112.10741
14. Piantadosi, S. T., & Aslin, R. (2016). Compositional reasoning in early childhood. PLOS ONE, 11(9), e0147734. https://doi.org/10.1371/journal.pone.0147734
15. Piantadosi, S. T., & Gallistel, C. R. (2022). A Foundation for Neuroscience. [Manuscript submitted for publication]. Faculty of Psychology, University of California, Berkeley.
16. Piantadosi, S. T., Palmeri, H., & Aslin, R. (2018). Limits on composition of conceptual operations in 9-month-olds. Infancy : The Official Journal of the International Society on Infant Studies, 23(3), 310–324. https://doi.org/10.1111/infa.12225
17. Pomiechowska, B., Brody, G., Teglas, E., & Kovacs, A. (2019). Twelve-month-olds use the principle of compositionality to combine newly learnt quantity labels with familiar kind labels.
18. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv. https://doi.org/10.48550/arXiv.2204.06125
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. arXiv. https://doi.org/10.48550/arXiv.2112.10752

20. Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7(4), 573–605. doi:10.1016/0010-0285(75)90024-9

21. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. arXiv. https://doi.org/10.48550/arXiv.2205.11487

22. Zhou, Y., & Lake, B. M. (2021). Flexible compositional learning of structured visual concepts. arXiv. https://doi.org/10.48550/arXiv.2105.09848