One-shot learning of generative speech concepts

Authors

Abstract

Keywords: speech recognition; category learning; one-shot learning; exemplar generation

Introduction

There has been recent interest in one-shot learning – the human ability to learn a new concept from just one or a few examples (e.g., Carey & Bartlett, 1978; Markman, 1989; Ahn, Brewer, & Mooney, 1992; Xu & Tenenbaum, 2007). Although one-shot learning is an important aspect of everyday cognition, traditional learning algorithms can require tens, hundreds, or thousands of examples before reaching a high level of classification performance. This mismatch poses a challenge to computational approaches for understanding human-level concept learning, yet over the last several decades, the fields of cognitive science and machine learning have made significant progress. Cognitive models have sought to quantitatively explain the way people generalize from just just a few examples in a low-dimensional space (Shepard, 1987; Feldman, 1997; Tenenbaum & Griffiths, 2001). Other cognitive models and computer vision algorithms become better one-shot learners through "transfer learning" or "learning to learn," where previous experience with related concepts helps to inform which dimensions or features are most important for generalization (Bart & Ullman, 2005; Colunga & Smith, 2005; Fei-Fei, Fergus, & Perona, 2006; Kemp, Perfors, & Tenenbaum, 2007).

Despite real progress spanning multiple disciplines, we are still far from a satisfying computational account of one-shot learning. Previous models have been limited by the simplicity of the representations that they learn – usually prototypes or exemplars in a feature space – which lack the power necessary for capturing many types of natural concepts (Murphy & Medin, 1985). While these feature-based approaches can be useful for classification, they provide less insight into how background knowledge interacts with learning or how people generalize in other ways beyond classification, including exemplar generation (Jern & Kemp, 2013), causal inference (Rehder, 2003), explanation (Williams & Lombrozo, 2010), and conceptual combination (Murphy, 1988). Given that people learn very rich concepts, even from just one or a few examples, the central computational challenge is to explain how people extract so much information from such limited data.

Analysis-by-synthesis is the classic idea, beginning with Helmholtz, that sensory data can be more richly represented by modeling the causal process that generates it. This has been an influential approach to studying perception in vision (Yuille & Kersten, 2006) and speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), where a perceptual classification decision can be made by selecting the category $c \in \{1, ..., K\}$, defined by parameters ψ_c, that maximizes the probability of generating a data example $x^{(i)}$

$$\underset{c}{\text{argmax}}\ P(x^{(i)}|\psi_c). \tag{1}$$

But learning is a more challenging problem, and there are good reasons to think an analysis-by-synthesis approach would not be successful for one-shot concept learning. Learning of a causal process from examples $x^{(1)}, ..., x^{(n)}$ can be formulated as the selection of a model ψ_c

$$\underset{\psi_c}{\text{argmax}} \prod_{i=1}^{n} P(x^{(i)}|\psi_c)P(\psi_c), \tag{2}$$

from a potentially infinite space of possible models. Since this can be a very difficult problem and can require many training examples to achieve good generalization (e.g., Geman, Bienenstock, & Doursat, 1992; Hinton & Nair, 2006), how could it possibly be learned from just a single training example? One-shot learning of generative models might only be sensible for a special subset of simple causal processes. If a human or machine was trying to learn what a "tree" was from just a single example, it would be hopeless to try to learn (or even represent) a fully-detailed process of biological growth, beginning with tree DNA and ending with the set of all possible trees. But at the right level of abstraction, the essence of a tree could be captured by a simple stochastic program, starting with a single branch and then recursively splitting until the tree terminates. With a prior $P(\psi_c)$ favoring simple generative processes, it seems possible to discover such a program from just a few examples.

What is the right notion of simplicity for formulating a prior over causal processes? Recent behavioral and computational work suggests that *compositionality* may be key principle for both encouraging representational simplicity and transferring previously learned knowledge from related concepts (Lake, Salakhutdinov, & Tenenbaum, 2012, 2013). Freely combining primitive structure can be a powerful way to build complex object representations (Biederman, 1987), and when combined with ideas from Hierarchical Bayesian modeling (Gelman, Carlin, Stern, & Rubin, 2004), it can also be a similarly powerful way to build a "generative model for generative models" that can perform one-shot learning. This idea was applied to the one-shot learning of handwritten character (Lake et al., 2013). Given a raw image of a new character, such as a character from a foreign alphabet, the model learns to represent it by a latent dynamic causal process, composed of pen strokes and their spatial relations (Fig. 1a). Sharing across characters is accomplished by the re-use of stochastic motor primitives (Fig. 1a-i) that can combine in new ways to make new characters (Fig. 1a-iv). From just a

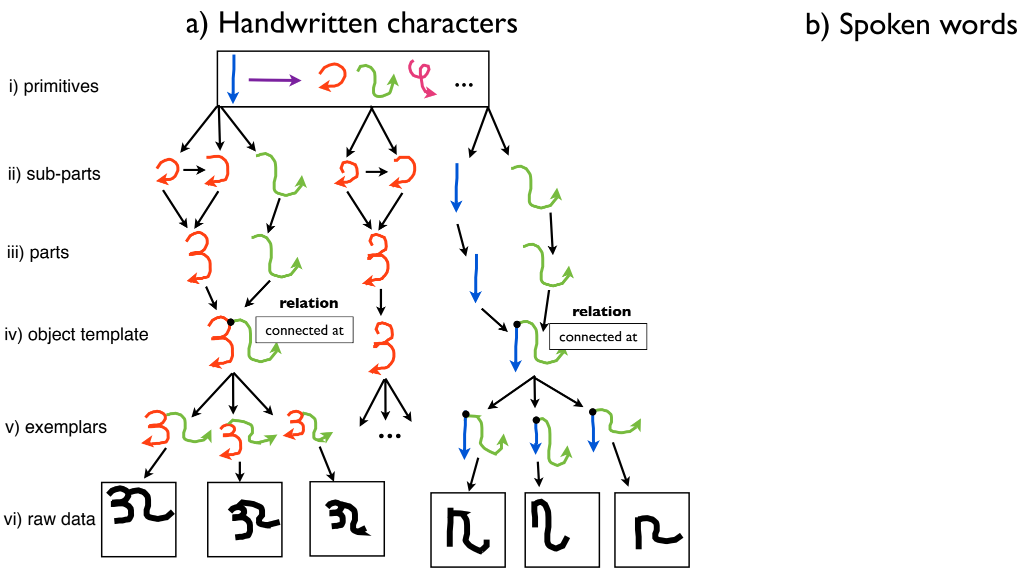


Figure 1: Hierarchical Bayesian modeling as applied to handwritten characters and speech.

single example, this model could both classify and generate new examples at a human level of performance.

How general is this approach, and could it apply to learning non-visual concepts? In this paper, we extend these ideas to the one-shot learning of new spoken words, such as a young child learning to recognize and pronounce a new word (the speech sound for "elephant") or an adult learning a word in a foreign language (like the word "elephant" in Japanese). Speech is a promising domain for our approach, since analysis-by-synthesis has been influential in speech recognition for decades (?, ?; Liberman et al., 1967) and there is a clear compositional structure of phonemes that could be exploited for learning. From the raw speech signal of a word, our model infers a causal representation based on a sequence of phone-like units (Fig. 1b). Like the characters, the prior distribution on these sequences are learned from experience with other words and can be combined together in new ways to define a new word (Fig. 1b-iv). Since each inferred word representation is itself generative, it can be used to both classify new examples or generate new examples (Fig. 1a-v) from just a single instance of a new word. By transferring a prior on sub-unit sequences learned from a corpus of Japanese speech, the model can classify new Japanese words at a level of accuracy similar to English-speaking humans. We also compare humans and the model on another natural form of generalization – an exemplar generation task.

- Related work in ASR on open dictionary words

Model

- Describe general ASR approach, and how this is related (but unsupervised, like a child)
- Mathematical description of the model

Experiment 1: Classification

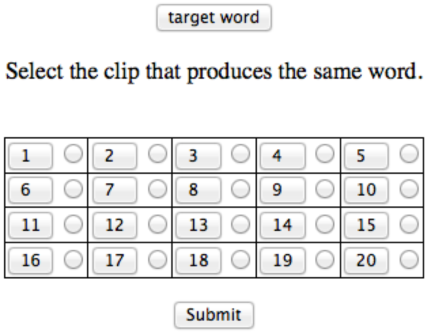
One-shot classification (20-way) using Japanese words. Here are factors we are going to manipulate.

People. Participants were recruited on Amazon's Mechanical Turk from a population of individual in the USA. All analyses were restricted to native English speakers that do not know any Japanese, as reported in a post-experiment survey rather than as a qualification for participation. This population was used for this and all other experiments.

Participants were asked to classify new Japanese words. To do so, they were shown a sequence of displays with a button labeled "target word" at the top of each display. Below it was a grid of buttons numbered "1" through "20," each with an associated radio button for response selection (and a general "submit" button to complete the trial). Participants were told that each button played a sound clip of a Japanese word, and their job was to pick the sound clip that produces the same word as the target word. Sound clips could be played more than once, and responses were not accepted until all buttons had been tried.

- Mention they never see the same word twice
- Explain word lengths, etc. here or in stimulus section
- explain instruction checks
- could click more than once, designed to minimize memory demands
- explain two conditions (same vs. different gender)

Hierarchical Bayesian model. Explain different model conditions and how it makes judgements.



Select the clip that produces the same word.

Figure 2: Screenshot of web interface for classification judgements. There was also a button to review the full instructions.

- Hierarchical Bayesian (HB) model (trained on English datasets)
- HB model (trained on Japanese, using the same corpus and speakers)

Dynamic Time Warp. This is a well-known approach in speech recognition for measure the similarity between two words that requires no learning (Sakoe & Chiba, 1978). Two sequences of feature frames are compared by computing an optimal non-linear warp, using dynamic programming, and then measuring the average distance between the frames of the sequence.

Results

Experiment 2: Generation

Classification is just one natural way in which people generalize from one example, and this experiment asked people and various models to generate new examples of a speech concept. Performance was measured by asking other participants to classify the generated examples, providing an indication of whether the new example belongs in the intended class. This is a weaker method for comparing humans and machines than a perceptual "Turing test," as used in our past work with characters (Lake et al., 2013). We did not use this stronger test because our model (and the field of speech synthesis more generally) is not yet up to the task, both in terms of emulating human voice and also in producing a range of different compelling examples instead of just one.

People. Ten participants were asked to say Japanese words after listening to a recording from a male voice. Each participant was assigned a different word length (3 through 12) and then completed twenty trials of speech recording using their computer's microphone. This procedure collected one sample per stimulus used in the previous experiment's classification task. One participant was replaced for very poor recording quality, and another was replaced for knowing some Japanese.

Hierarchical Bayesian model.

- full model (English vs. Japanese)
- 25% noise
- 50% nosie
- no primitives

Evaluation procedure. Using a within-subjects design, thirty participants classified a mix of synthesized examples from both people and the comparison models. The trials appeared as they did in Experiment 1 (Fig. 2) where the "target word" button played a synthesized example. The option clips played original Japanese recordings, matched for word length within a trial as in Experiment 1. Since the synthesized examples were based on male clips, only the female clips were used as options. There was one practice trial (in English) followed by 50 trials with the synthesized example drawn uniformly from the set of all synthesized samples. Since the example sounds vary in quality and some are hardly speech-like, participants were told that the sound quality varies, may be very poor, or may sound machine generated but they were encouraged to try their best. Also, the instructions and practice trial were changed to include a degraded target word clip. All clips were normalized for volume.

Results

Several people commented about the task being too long and too difficult, and two participants were removed for guessing.[1] The main pattern of results are shown in Fig. 3.

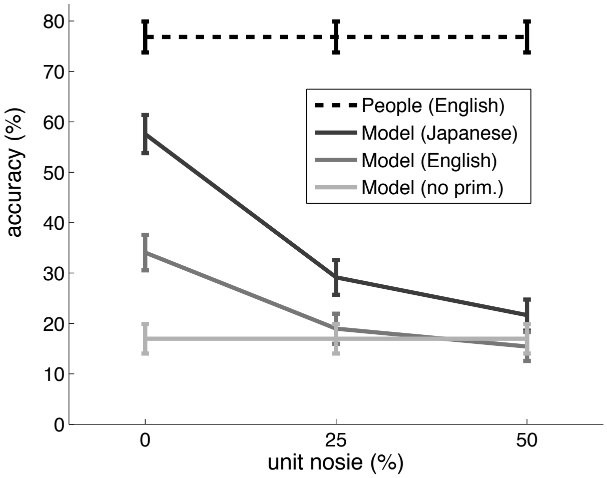


Figure 3: Percent of synthesized examples that were identified as belonging to the correct category.

Replication

Explain the between subjects replication that we ran here.

[1] Participants spent between 19 minutes to 87 minutes on the task, and there was also a significant correlation between overall accuracy and time spent (R=0.58, $p<0.001$). In a conservative attempt to eliminate guessing, two participants were removed for failing to listen to the "target word" at least twice on per trial on average (6 times was the experiment average). This made little difference for the pattern of results.

Discussion

References

Ahn, W.-k., Brewer, W. F., & Mooney, R. J. (1992). Schema Acquisition From a Single Example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 391–412.

Bart, E., & Ullman, S. (2005). Cross-generalization: Learning novel classes from a single example by feature replacement. In *Computer vision and pattern recognition (cvpr).*

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115–47.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Colunga, E., & Smith, L. B. (2005, April). From the lexicon to expectations about kinds: a role for associative learning. *Psychological review*, *112*(2), 347–82. doi: 10.1037/0033-295X.112.2.347

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*, 145–170.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). Chapman and Hall/CRC.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*, 1–58.

Hinton, G. E., & Nair, V. (2006). Inferring motor programs from images of handwritten digits. In *Advances in Neural Information Processing Systems 19.*

Jern, A., & Kemp, C. (2013, March). A probabilistic account of exemplar and category generation. *Cognitive psychology*, *66*(1), 85–125. doi: 10.1016/j.cogpsych.2012.09.003

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2012). Concept learning as motor program induction: A large-scale empirical study. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2013). One-shot learning by inverting a compositional causal process. In *Advances in neural information processing systems 26.*

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Markman, E. M. (1989). *Categorization and Naming in Children*. Cambridge, MA: MIT Press.

Murphy, G. L. (1988, December). Comprehending complex concepts. *Cognitive Science*, *12*(4), 529–562. doi: 10.1016/0364-0213(88)90012-2

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316. doi: 10.1037/0033-295X.92.3.289

Rehder, B. (2003, November). A causal-model theory of conceptual representation and categorization. *Journal of experimental psychology. Learning, memory, and cognition*, *29*(6), 1141–59. doi: 10.1037/0278-7393.29.6.1141

Sakoe, H., & Chiba, S. (1978, February). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. doi: 10.1109/TASSP.1978.1163055

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–40.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive science*, *34*(5), 776–806.

Xu, F., & Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*(2), 245–272.

Yuille, A., & Kersten, D. (2006, July). Vision as Bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, *10*(7), 301–8. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/16784882` doi: 10.1016/j.tics.2006.05.002