Continual adaptation for efficient machine communication

Robert D. Hawkins [*1] **Minae Kwon** [*2] **Dorsa Sadigh** [2] **Noah D. Goodman** [1 2]

Abstract

To communicate with new partners in new contexts, humans rapidly form new linguistic conventions. Recent language models trained with deep neural networks are able to comprehend and produce the existing conventions present in their training data, but are not able to flexibly and interactively adapt those conventions on the fly as humans do. We introduce a repeated reference task as a benchmark for models of adaptation in communication and propose a regularized continual learning framework that allows an artificial agent initialized with a generic language model to more accurately and efficiently understand their partner over time. We evaluate this framework through simulations on COCO and in real-time reference game experiments with human partners.

1. Introduction

Linguistic communication depends critically on shared knowledge about the meanings of words (Lewis, 1969). However, the real-world demands of communication often require speakers and listeners to go *beyond* dictionary meanings to understand one another (Clark, 1996; Stolk et al., 2016). The social world continually presents new communicative challenges, and agents must continually coordinate on new meanings to meet them.

For example, consider a nurse visiting a bed-ridden patient in a cluttered home. The first time they ask the nurse to retrieve a particular medication, the patient must painstakingly refer to unfamiliar pills, e.g. "the vasoprex-tecnoblek meds for my blood pressure, in a small bluish bottle, on the bookcase in my bathroom." After a week of care, however, they may just ask for their "Vasotec."

This type of flexible language use poses a challenge for models of language in machine learning. Approaches based on deep neural networks typically learn a monolithic meaning function during training, with fixed weights during use. For an in-home robot to communicate as flexibly and efficiently with patients as a human nurse, it must be equipped with a continual learning mechanism. Such a mechanism would present two specific advantages for interaction and communication applications. First, to the extent that current models have difficulty communicating in a new setting, an adaptive approach can quickly improve performance on the relevant subset of language. Second, for human-robot contexts, an adaptive model enables speakers to communicate more *efficiently* as they build up common ground, remaining understandable while expending significantly fewer words as humans naturally do (Clark & Wilkes-Gibbs, 1986).

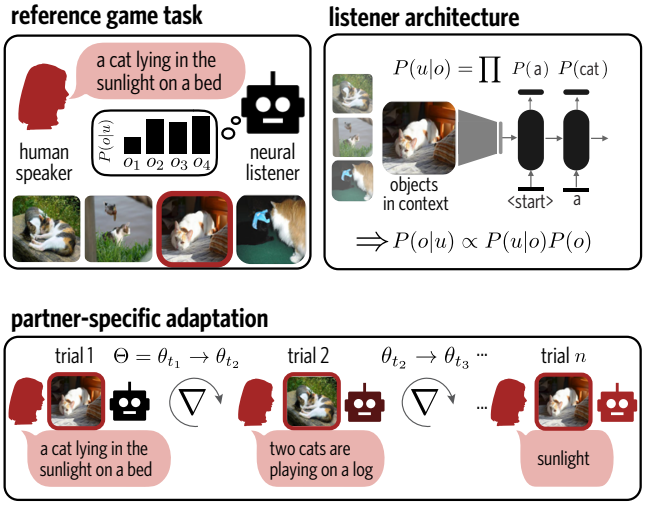

Figure 1. Task, architecture, and continual learning approach.

In this paper, we introduce a benchmark communication task and general continual learning framework for transforming neural language models into *adaptive* models that can be deployed in real-time interactions with other agents.

Our key insight is that through continual interactions with the same partner in a shared context, an adaptive listener can more effectively communicate with its partner (Fig. 1).

We are motivated by hierarchical Bayesian approaches to task-specific adaptation. Our approach integrates two core components: (i) a loss function combining speaker and listener information, and (ii) a regularization scheme for fine-tuning model weights without overfitting.

[*]Equal contribution [1]Department of Psychology, Stanford University, Stanford, CA [2]Department of Computer Science, Stanford University, Stanford, CA. Correspondence to: Robert Hawkins <rxdh@stanford.edu>.

Proceedings of the 1st Adaptive & Multitask Learning Workshop, Long Beach, California, 2019.

2. Approach

We begin by recasting communication as a multi-task problem for meta-learning. Each context and communicative partner can be regarded as a related but distinct task making its own demands on the agent's language model. To be effective across many such tasks, a communicative agent must both (1) have a good prior representation they can use to understand novel partners and contexts, and (2) have a mechanism to rapidly update this representation from a small number of interactions.

2.1. Repeated reference game task

As a benchmark for studying this problem, we introduce the *repeated reference game* task (Fig. 1), which has been widely used in cognitive science to study partner-specific adaptation in communication (Krauss & Weinheimer, 1964; Clark & Wilkes-Gibbs, 1986; Wilkes-Gibbs & Clark, 1992). In this task, a speaker agent and a listener agent are shown a context of images, $\mathcal{C}$, and must collaborate on how to refer to them. On each trial, one of these images is privately designated as the *target object* o for the speaker. The speaker thus takes the pair $(o, \mathcal{C})$ as input and returns an utterance u that will allow the listener to select the target. The listener agent then takes $(u, \mathcal{C})$ as input and returns a softmax probability for each image, which it uses to make a selection. Both agents then receive feedback about the listener's response and the identity of the target. Critically, the sequence of trials is constructed so that each image repeatedly appears as the target, allowing us to evaluate how communication about each image changes over time.

2.2. Continual adaptation with Hierarchical Bayes

Before formalizing our algorithm as a generic update rule for neural networks, we describe the theoretical Bayesian foundations of our approach. At the core of any communication model is a notion of the *semantics* of language, which supplies the relationship between utterances and states of the world. Under a Bayesian approach, this representation is probabilistic: we represent some uncertainty over meanings. In a hierarchical Bayesian model, this uncertainty is structured over different partners and contexts.

At the highest level of the hierarchy is a *task-general* variable Θ which parameterizes the agent's task-specific prior expectations $P(\theta_i|\Theta)$, where θ_i represents the semantics used by a novel partner i. Given observations D_i from communicative interactions in that context, an agent can update their *task-specific* model using Bayes rule:

$$P(\theta_i|D_i, \Theta) \propto P(D_i|\theta_i)P(\theta_i|\Theta) \tag{1}$$

The Bayesian formulation thus decomposes the problem of task-specific adaptation into two terms, a prior term $P(\theta_i|\Theta)$ and a likelihood term $P(D_i|\theta_i)$. The prior captures the idea that different language tasks share some task-general structure in common: in the absence of strong information about usage departing from this common structure, the agent ought to be regularized toward their task-general knowledge.

The likelihood term accounts for needed deviations from general knowledge due to evidence from the current situation. The form of the likelihood depends on the task at hand. For our benchmark communication task, $D_i = \{(u, o)_t\}$ contains paired observations of utterances u and their objects of reference o at times t. These data can be viewed from the point of view of a speaker (generating u given o) or a listener (choosing o from a context of options, given u) (Smith et al., 2013; Hawkins et al., 2017). A *speaker* model[1] uses its task-specific semantics θ_i to sample utterances u proportional to how well they apply to o:

$$P_S(u|o, \theta_i) \propto \exp f_{\theta_i}(u, o) \tag{2}$$

A *listener* can be modeled as inverting this speaker model to evaluate how well an utterance u describes each object o *relative to the others* in a context $\mathcal{C}$ of objects by normalizing (Frank & Goodman, 2012; Vedantam et al., 2017; Cohn-Gordon et al., 2018; Monroe et al., 2017):

$$P_L(o|u, \mathcal{C}, \theta_i) \propto P_S(u|o, \theta_i)P(o) \tag{3}$$

Because these views of the data D_i provide complementary statistical information about the task-specific semantics θ_i, we will combine them in our loss.

2.3. Continual adaptation for neural language models

There is a deep theoretical connection between the hierarchical Bayesian framework presented in the previous section and recent deep learning approaches to multi-task learning (Nagabandi et al., 2018; Grant et al., 2018; Jerfel et al., 2018). Given a task-general initialization, regularized gradient descent on a particular task is equivalent to conditioning on new data under a Bayesian prior. We exploit this connection to propose an online continual learning scheme for a neural listener model that can adapt to a human speaker in our challenging referential communication task.

Concretely, we consider an image-captioning network that combines a convolutional visual encoder (ResNet-152) with an LSTM decoder (Vinyals et al., 2015). The LSTM takes a 300-dimensional embedding as input for each word in an utterance and its output is then linearly projected back to a softmax distribution over the vocabulary size. To pass the visual feature vector computed by the encoder into the decoder, we replaced the final layer of ResNet with a fully-connected adapter layer. This layer was jointly pre-trained with the decoder on the COCO training set and then frozen,

[1]The function f abstracts away from any specific architecture.

Algorithm 1 Update step for adaptive language model

Input: θ_t: weights at time t
Output: θ_{t+1}: updated weights
Data: (u_t, o_t): observed utterance and object at time t
for step **do**
  sample augmented batch of sub-utterances $\mathbf{u} \sim \mathcal{P}(u)$
  update $\theta_t \leftarrow \theta_t + \beta\nabla[P(\mathbf{u}|o) + P(o|\mathbf{u}) + \text{reg}(o, u)]$
end for

leaving only the decoder weights (i.e. word embeddings, LSTM, and linear output layer) to be learned in an online fashion. Upon observing each utterance-object data point in the current task, we take a small number of gradient steps fine-tuning these weights to better account for the speaker's usage (see Algorithm 1). We consider several loss terms and techniques to do so.

Speaker and listener likelihood. The primary signal available for adaptation is the (log-) probability of the new data under speaker and listener likelihoods given in Eqns. 2-3. Our speaker likelihood serves to make the observed utterance more likely for the target in *isolation*, while our listener likelihood makes it more likely *relative* to other objects in context. The speaker and listener likelihoods can be computed directly from the neural captioning model, as shown in Fig. 1, where the probability of each word is given by the softmax decoder output conditioned on the sentence so far.

Regularization. We introduce two kinds of regularization terms to approximate the Bayesian prior on task-specific learning. First, rather than directly regularizing weights, a *global KL regularization* term minimizes the divergence between the captioning model's output probabilities before and after fine-tuning (Yu et al., 2013; Galashov et al., 2018). Since the support for our distribution of captions is infinite, we approximate the divergence incrementally by expanding from the maximum a posteriori (MAP) word at each step according to P, where P represents the model at initialization and Q_t represents the model at time t. This loss is then averaged across random images from the full domain $\mathcal{O}$, not just those in context:

$$\sum_{o\in\mathcal{O}}\sum_{i<L} D_{\text{KL}}\left(P(w_i|o, w_{i-1}^{\text{MAP}})||Q_t(w_i|o, w_{i-1}^{\text{MAP}})\right) \quad (4)$$

where we denote the word at position i by w_i and terminate after reaching L, the length of the MAP caption. A second form of regularization we consider is *local rehearsal*: we sum the likelihood over previous observations $(u, o)_\tau$ from the same partner to prevent overfitting to the most recent observation. Finally, we examine *listener* variants of both forms of regularization by using the listener likelihood instead of the speaker likelihood. For example, we compute the listener KL regularization by comparing the initial listener distribution over the objects in context $o \in \mathcal{C}$ with the fine-tuned model's distribution: $D_{\text{KL}}\left(P(o|u)||Q_t(o|u)\right)$. We anneal the weight on the listener regularization terms over time while reverse-annealing the listener likelihood.

Data augmentation. A final component of our approach is a data augmentation step on the new utterance u. Ideally, an adaptive agent should learn that words and sub-phrases contained in the observed utterance are compositionally responsible for its meaning. We thus derive a small training dataset $D(u)$ from u; for simplicity, we take the (ordered) powerset $D(u) = \mathcal{P}(u)$ of all sub-utterances.[2]

3. Evaluations

To evaluate our model, we implemented a repeated reference game using images from the validation set of COCO (Lin et al., 2014) as the targets of reference. To construct challenging contexts $\mathcal{C}$, we used our pre-trained visual encoder to find sets of highly similar images. We extracted feature vectors for each image, partitioned the images into 100 groups using a k-means algorithm, sampled one image from each cluster, and took its 3 nearest neighbors in feature space, yielding 100 unique contexts of 4 images each[3].

3.1. Human baselines

We first investigated the baseline performance of human speakers and listeners. We recruited 113 participants from Amazon Mechanical Turk and automatically paired them into an interactive environment with a chatbox. For each of these 56 pairs, we sampled a context and constructed a sequence of 24 trials structured into 6 repetition blocks, where each of the 4 images appeared as the target once per block. We prevented the same target appearing twice in a row and scrambled the order of the images on each player's screen on each trial.

We found that pairs of humans were remarkably accurate at this task, with performance near ceiling on every round. At the same time, they grew increasingly efficient in their communication: the utterance length decreased from an average of 7 words per image on the first repetition to only 3 words on the last. A mixed-effects regression with random slopes and intercepts accounting for variability at the pair- and context-level found a significant decrease in utterance length across repetitions, $t = -5.8, p < 0.001$ (Fig. 2A).

3.2. Model evaluation with human partner

Next, we evaluated how our adaptive listener performed in *real-time interaction* with human speakers. We recruited 45 additional participants from Amazon Mechanical Turk

[2]Grammatical acceptability could in principle be taken into account using alternative sets derived from a syntactic parse.

[3]Using pre-trained VGG as the encoder gave similar contexts.

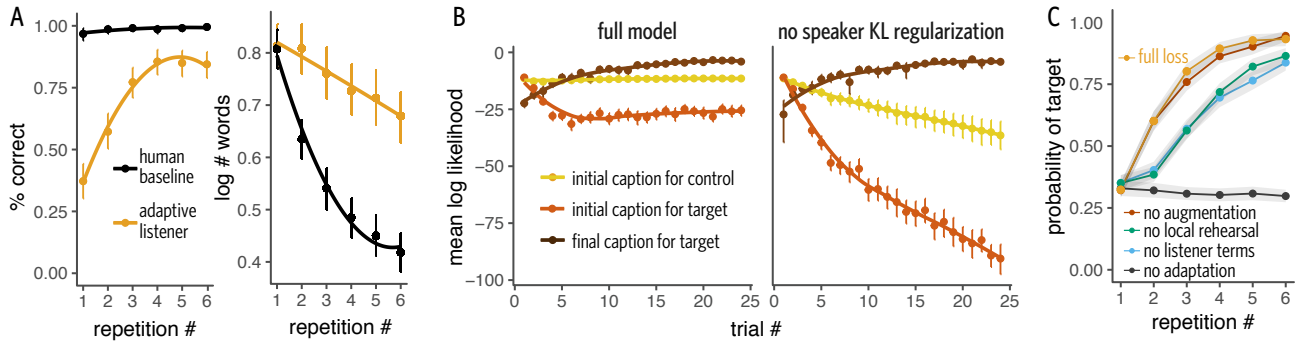

Figure 2. (A) Human speakers grow more efficient and accurate as our model adapts. Curves show regression fits. (B) Speaker KL regularization prevents catastrophic forgetting. (C) Lesions reveal the contributions of each loss term. Error bars and ribbons are bootstrapped 95% CIs.

who were told they would be paired with an artificial agent learning how they talk. This task was identical to the one performed by humans, except participants were only allowed to enter a single message through the chatbox on each trial. This message was sent to a GPU where the model weights from the previous trial were loaded, used to generate a response, and updated in real-time for the next round. The approximate latency for the model to respond was 6-8s.

We used a batch size of 8, learning rate of 0.0005, and took 8 gradient steps after each trial. For our loss objective, we used a linear combination of all speaker and listener likelihood losses and regularization terms. We found that a listener based on a pre-trained neural captioning model—the initialization for our adapting model—performs much less accurately than humans due to the challenging nature of the reference task. Yet our model rapidly improves in accuracy as it coordinates on appropriate meanings with human speakers. Similarly, while speakers did not simplify their utterances to the same extent as they did with other humans, perhaps due to early feedback about errors, they nonetheless became significantly more efficient over time, $b = -19, t = -5$ (see Fig. 2A).

4. Analysis

We proceed to a series of lesion analyses that analyze the role played by each component of our approach.

4.1. Preventing catastrophic forgetting

Fine-tuning repeatedly on a small number of data points presents a clear risk of catastrophic forgetting (Robins, 1995), losing our ability to produce or understand utterances for other images. Our KL regularization term (Eqn. 4) was intended to play the same role as a Bayesian prior, preventing catastrophic forgetting by tethering task-specific behavior to the task-general model. To test the effectiveness of this term, we examined the likelihood of different captions before and after adaptation to the human baseline utterances. First, we sampled a random set of images from COCO that were not used in our experiment as *control* images, and used the initialized state of the LSTM to greedily generate a caption for each. We also generated initial captions for the *target* objects in context. We recorded the likelihood of all of these sampled captions under the model at the beginning and at each step of adaptation until the final round. Finally, we greedily generated an utterance for each target at the end and retrospectively evaluated its likelihood at earlier states. These likelihood curves are shown with and without speaker KL regularization in Fig. 2B. The final caption becomes more likely in both cases; without the KL term, the initial captions for both targets and unrelated controls are (catastrophically) lost.

4.2. Lesioning loss terms

We next simulated our adaptive agent's performance understanding utterances from the human baseline under lesioned losses (Fig. 2C). We found that rehearsal on previous rounds had the largest qualitative benefit, allowing for faster adaptation on early rounds, while data augmentation and the listener terms provided small boosts later in the game. Compared to a non-adapting baseline, however, even a simple loss only containing the speaker likelihood and speaker KL regularization performed better over time—successfully adapting to human language use.

5. Conclusions

Human language use is flexible, continuously adapting to the needs of the current situation. In this paper, we introduced a challenging repeated reference game benchmark for artificial agents, which requires such adaptability to succeed. We proposed a continual learning approach that forms context-specific conventions by adapting general-purpose semantic knowledge. Even when models based on general-purpose knowledge perform poorly, our approach allows human speakers working with adapted variants of such models to become more accurate and more efficient over time.

References

Clark, H. H. *Using language*. Cambridge university press, 1996.

Clark, H. H. and Wilkes-Gibbs, D. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

Cohn-Gordon, R., Goodman, N., and Potts, C. Pragmatically Informative Image Captioning with Character-Level Reference. *arXiv preprint arXiv:1804.05417*, 2018.

Frank, M. C. and Goodman, N. D. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Galashov, A., Jayakumar, S. M., Hasenclever, L., Tirumala, D., Schwarz, J., Desjardins, G., Czarnecki, W. M., Teh, Y. W., Pascanu, R., and Heess, N. Information asymmetry in KL-regularized RL. In *ICLR*, 2018.

Grant, E., Finn, C., Levine, S., Darrell, T., and Griffiths, T. Recasting Gradient-Based Meta-Learning as Hierarchical Bayes. *arXiv preprint arXiv:1801.08930*, 2018.

Hawkins, R. X. D., Frank, M. C., and Goodman, N. D. Convention-formation in iterated reference games. In *Proceedings of the 39th annual meeting of the cognitive science society*, 2017.

Jerfel, G., Grant, E., Griffiths, T. L., and Heller, K. Online gradient-based mixtures for transfer modulation in meta-learning. *arXiv:1812.06080*, 2018.

Krauss, R. M. and Weinheimer, S. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1-12):113–114, 1964.

Lewis, D. *Convention: A philosophical study*. Harvard University Press, 1969.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Monroe, W., Hawkins, R. X. D., Goodman, N. D., and Potts, C. Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *arXiv preprint arXiv:1703.10186*, 2017.

Nagabandi, A., Finn, C., and Levine, S. Deep Online Learning via Meta-Learning: Continual Adaptation for Model-Based RL. *arXiv:1812.07671*, 2018.

Robins, A. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

Smith, N. J., Goodman, N., and Frank, M. Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems*, pp. 3039–3047, 2013.

Stolk, A., Verhagen, L., and Toni, I. Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, 20(3):180–191, 2016.

Vedantam, R., Bengio, S., Murphy, K., Parikh, D., and Chechik, G. Context-aware Captions from Context-agnostic Supervision. *arXiv preprint arXiv:1701.02870*, 2017.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and Tell: A Neural Image Caption Generator. In *CVPR*, pp. 3156–3164, 2015.

Wilkes-Gibbs, D. and Clark, H. H. Coordinating beliefs in conversation. *Journal of memory and language*, 31(2): 183–194, 1992.

Yu, D., Yao, K., Su, H., Li, G., and Seide, F. Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7893–7897. IEEE, 2013.