Supplemental Materials for Continual Adaptation for Efficient Machine Communication

Anonymous EMNLP submission

Appendix A: derivation of incremental KL

We denote the distribution over a sequence of T tokens by $p(w_{1:T}) = p(w_1, w_2, \ldots, w_T)$. We are interested in the KL divergence between two such distributions, $KL\left(p(w_{1:T}) \,\|\, q(w_{1:T})\right)$. We show that our approximation over possible captions is the best incremental estimator of this intractable objective. First, note that the KL divergence factors in the following way.

Lemma 1.

$$\begin{aligned} &KL\left(p(w_{1:2}) \,\|\, q(w_{1:2})\right) \\ &\quad = KL\left(p(w_1) \,\|\, q(w_1)\right) \\ &\qquad + \mathbb{E}_{p(w_1)} KL\left(p(w_2|w_1) \,\|\, q(w_2|w_1)\right) \end{aligned}$$

Proof.

$$\begin{aligned} &KL\left(p(w_{1:2}) \,\|\, q(w_{1:2})\right) \\ &= \sum_{w_1}\sum_{w_2} p(w_{1:2}) \log \frac{p(w_{1:2})}{q(w_{1:2})} \\ &= \sum_{w_1}\sum_{w_2} p(w_{1:2}) \log \frac{p(w_1)}{q(w_1)} \\ &\quad + \sum_{w_1}\sum_{w_2} p(w_{1:2}) \log \frac{p(w_2|w_1)}{q(w_2|w_1)} \\ &= \sum_{w_1} \log \frac{p(w_1)}{q(w_1)} \sum_{w_2} p(w_{1:2}) \\ &\quad + \sum_{w_1} p(w_1) \sum_{w_2} p(w_2|w_1) \log \frac{p(w_2|w_1)}{q(w_2|w_1)} \\ &= KL\left(p(w_1) \,\|\, q(w_1)\right) \\ &\quad + \mathbb{E}_{p(w_1)} KL\left(p(w_2|w_1) \,\|\, q(w_2|w_1)\right) \end{aligned}$$

Now, let w_1^* be the token at which $p(w_1)$ takes its maximum value. Then w_1^* is the best single-sample approximation of the expectation:

$$\begin{aligned} &\mathbb{E}_{p(w_1)} KL\left(p(w_2|w_1) \,\|\, q(w_2|w_1)\right) \\ &\quad \approx KL\left(p(w_2|w_1^*) \,\|\, q(w_2|w_1^*)\right) \end{aligned}$$

If we assume that $p(w_{1:T})$ is Markov (as in a recurrent model) then it follows from repeatedly applying the lemma that

$$\begin{aligned} &KL\left(p(w_{1:T}) \,\|\, q(w_{1:T})\right) \\ &= \sum_{i=1}^{T} KL\left(p(w_i|w_1^*, \ldots, w_{i-1}^*) \,\|\, q(w_i|w_1^* \ldots, w_{i-1}^*)\right) \\ &= \sum_{i=1}^{T} KL\left(p(w_i|w_{i-1}^*) \,\|\, q(w_i|w_{i-1}^*)\right) \end{aligned}$$

recovering our objective.

Appendix B: Parameter settings

For both the speaker task and listener task, we used a learning rate of 0.0005, took 6 gradient steps after each trial, and used a batch size of 8 when sampling utterances from the augmented set of sub-phrases, At each gradient step, we sampled 50 objects from the full domain $\mathcal{O}$ of COCO to approximate our regularization term. We set the coefficients weighting each term in our loss function as follows: 1.0 (utterance loss), 0.1 (contrastive loss), 0.5 (KL regularization), 0.3 (local rehearsal).

Appendix C: Regression details

To formally test increasess in efficiency reported for baseline pairs of humans in Sec. 4.1 (see Fig. 3 and S2), we conducted a mixed-effects regression predicting utterance length. We included a fixed effect for context type (i.e. 'simple' vs. 'challenging') as well as orthogonalized linear and quadratic effects of repetition number, and each of their interactions with context type. We also included random intercepts accounting for variability in initial utterance length at the pair- and image-level. To test increases in the adaptive listener model's accuracy in Sec. 4.2, we conducted a mixed-effects logistic regression on trial-level responses (i.e. 'correct'

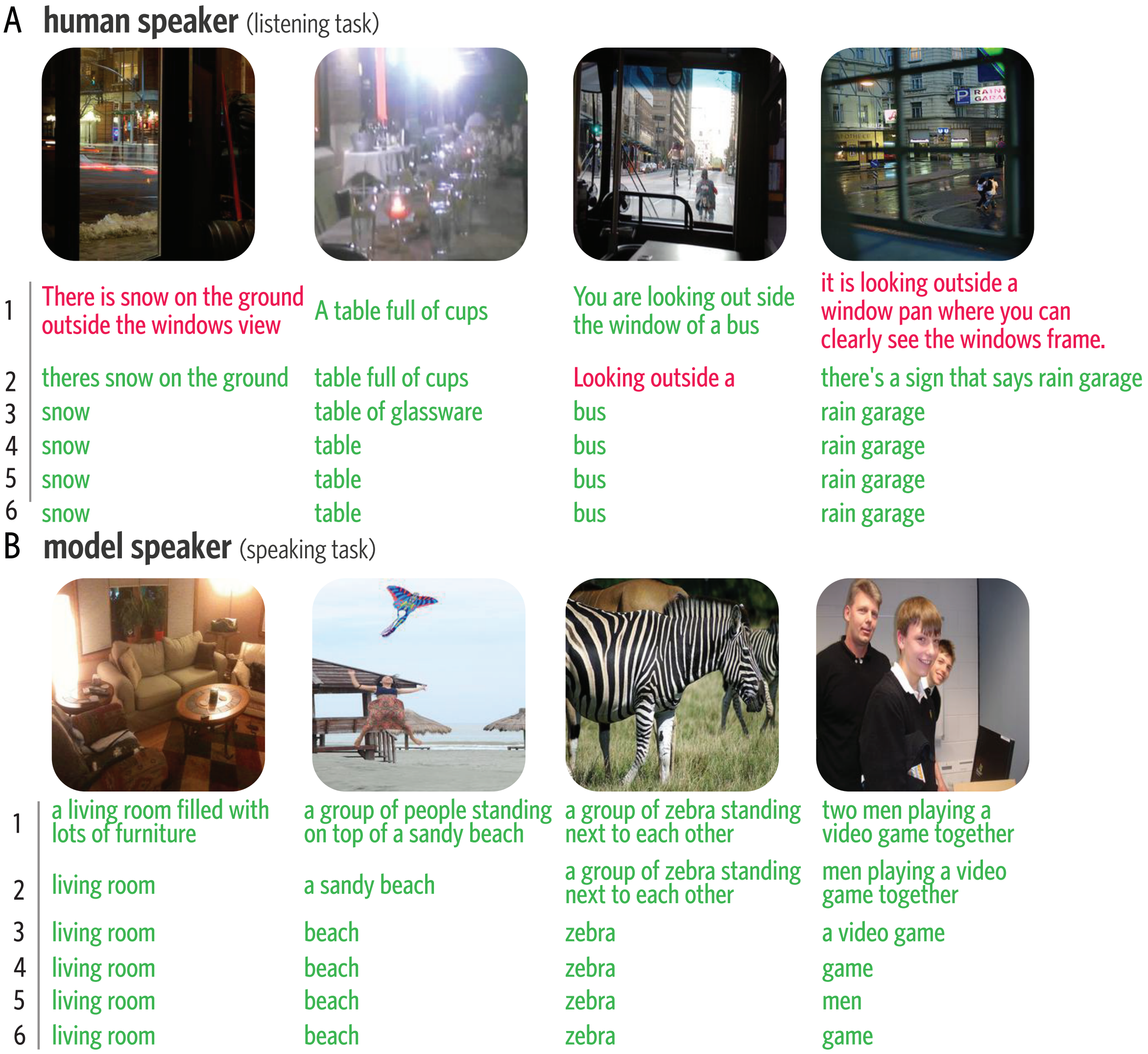
A human speaker (listening task)
1
There is snow on the ground outside the windows view
A table full of cups
You are looking out side the window of a bus
it is looking outside a window pan where you can clearly see the windows frame.
2
theres snow on the ground
table full of cups
Looking outside a
there's a sign that says rain garage
3
snow
table of glassware
bus
rain garage
4
snow
table
bus
rain garage
5
snow
table
bus
rain garage
6
snow
table
bus
rain garage
B model speaker (speaking task)
1
a living room filled with lots of furniture
a group of people standing on top of a sandy beach
a group of zebra standing next to each other
two men playing a video game together
2
living room
a sandy beach
a group of zebra standing next to each other
men playing a video game together
3
living room
beach
zebra
a video game
4
living room
beach
zebra
game
5
living room
beach
zebra
men
6
living room
beach
zebra
game

Figure S1: Complete set of referring expressions produced by (A) a *human speaker* interacting with our listener model and (B) our *speaker model* interacting with a human partner, as described below in Appendix D. Utterances are color-coded with the response accuracy. Green is correct; red is incorrect.

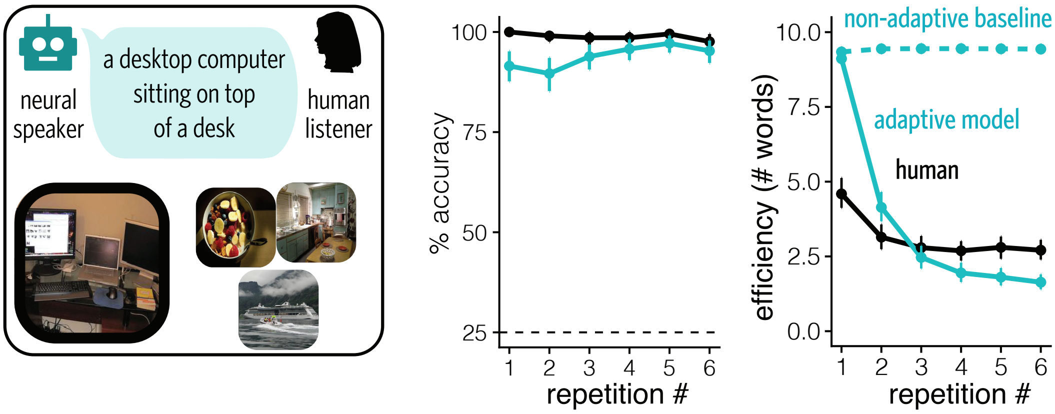

Figure S2: Speaker model evaluations with human listener. Error ribbons are bootstrapped 95% CIs.

vs. 'incorrect') with the same effect structure, removing context type, as this experiment was only conducted on challenging contexts. This same effect structure was used to test improvements in the adaptive speaker model's efficiency in Sec. 4.3.

We tested our listener ablations in Sec. 5.2 using a mixed-effects logistic regression with fixed effects of repetition number and model variant, as well as participant-level random intercepts and slopes (for repetition number). We dummy-coded the models setting the baseline at the full model, to facilitate direct comparison.

Appendix D: Speaker evaluation and analysis

Task description

We designed the speaking task such that the pre-trained model's *efficiency* — the number of words needed to identify the target — would be poor at the outset. Because the COCO captions seen during pre-training were relatively exhaustive (i.e. mentioning many attributes of the image), we required *simple* contexts where the pre-trained model would produce more complex referring expressions than required to distinguish the images. To construct simple contexts we sampled images randomly from different COCO category labels. For example, one context might contain an image of an elephant, an image of a boat, and so on.

Evaluation results

We evaluated our model in the *speaking* task using simple contexts, which requires the model to form more efficient conventions given feedback from human responses. 53 participants from Amazon Mechanical Turk were paired to play the listener role with our speaker model. Utterances were selected from the LSTM decoder using beam search with a beam width of 50 and standard length normalization to mitigate the default bias against long utterances (e.g. Wu et al., 2016). After producing an utterance, the model received feedback about the listener's selection. If its partner correctly selected the intended target, it proceeded to adapt conditioning on the new observation; in the event of an incorrect response, it refrained from updating. This strategy thus only leads to inferences about utterance meanings (and sub-phrase meanings, through data augmentation) after positive evidence of understanding.

As expected, the model starts with much longer captions than human speakers use in simple contexts (Fig. S2). It uses nearly as many words for simple contexts as humans used for challenging contexts. However, it gets dramatically more efficient over interaction while maintaining high accuracy. We found a significant decrease in utterance length over successive repetitions, $t = 35$, $p < 0.001$, using the same mixed-effects regression structure reported above. A non-adapting baseline shows no improvement, as it has no mechanism for changing its expectations about utterances over time.

Pragmatic reasoning supports speaker informativity

We now proceed to analyze the *speaking task*, beginning with the role of pragmatic reasoning. In principle, incorporating pragmatic reasoning during adaptation (i.e. in our contrastive likelihood term) introduces an inductive bias for *mutual exclu-*

Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5	Rep. 6
a group of people standing on a sandy beach	a group of people standing on top	a group of people standing	a group of people	a group of	a group
a couple of zebra standing next to trees	a couple of zebra standing next	a couple of zebra	a couple of	a couple	a couple
a living room filled with lots of furniture	a living room filled with lots furniture	a living room filled with furniture	a living room	a living	a living

Table S1: Examples of utterances produced by ablated speaker with pure cost penalty instead of data augmentation, which quickly become ungrammatical and incoherent.

sivity (Smith et al., 2013; Frank et al., 2009; Gandhi and Lake, 2019). When the listener correctly selects the target, the speaker not only learns that the listener believes this is a good description for the target but can also infer that the listener *does not* think it is a good description for the other objects in context; otherwise, they would have selected one of the other objects instead. Thus, in addition to boosting listener adaptation, we expected explicit pragmatic reasoning to allow the speaker to gradually produce more informative, distinguishing utterances.

The *challenging* contexts provide an ideal setting for evaluating speaker pragmatics, because the pretrained speaker model initially produces the same caption for all four images in context. We simulated games with our adaptive speaker as well as an ablated variant with no contrastive term in its adaptation objective. The model was always given feedback that the correct target was selected. We measured informativity by examining the proportion of words that overlapped between the utterances produced for the different images in a particular context: $|u_i \cap u_j| / \min(|u_i|, |u_j|)$ for combinations of utterances (u_i, u_j) where $i \neq j$. This measure ranges between 0% when the intersection is empty and 100% in the event of total overlap.

Even though both the full model and the ablated variant initially produce completely overlapping utterances, we found a rapid differentiation of utterances for the model with pragmatics intact, as each image becomes associated with a distinct label. Meanwhile, the ablated version continues to have high overlap even on later repetitions: it often ends up producing the same one-word label for multiple objects as it reduces (Fig. S3A).

Compositional data augmentation supports efficiency

Finally, we investigated the role played by the compositional data augmentation mechanism for allowing our speaker model to become more efficient (Fig. S3B). Two concerns may be raised about this mechanism. First, it is possible that the RNN decoder architecture is already able to appropriately update expectations about sub-parts from the whole without being given explicit parses, so augmentation is redundant. Second, it may be argued that this augmentation mechanism just imposes a glorified length penalty *forcing* the speaker to shorten, rather than allowing efficiency to come out of the model naturally.

To address these concerns, we compare augmentation with two variants: (1) an ablated model with no augmentation, and (2) an alternative mechanism that explicitly imposes a length cost at production time. This alternative is implemented by re-ranking the top 25 utterances from beam search according to $U(u_i) = P(u_i)/\ell(w) - \beta_w \ell(w)$ where the first term is the length-normalized beam-search objective and the second term is an explicit bias for shorter utterances. When $\beta_w = 0$, this is equivalent to top-k beam search but as $\beta_k \to \infty$, the model will increasingly prefer short utterances.

We simulated the behavior of these model variants in each of the 53 games we collected in our

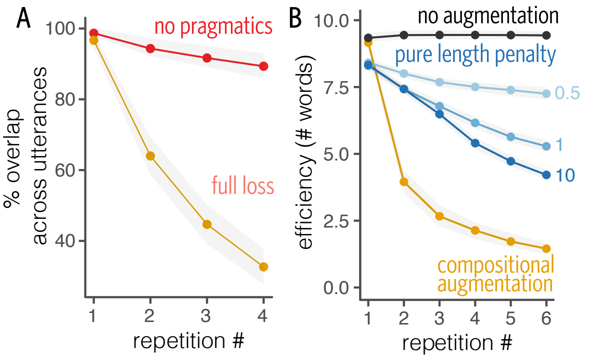

Figure S3: Speaker model ablations. (A) The contrastive loss allows the model to become informative in challenging contexts. (B) Compositionally augmenting adaptation data with sub-phrases of the utterance allows stronger gains in efficiency than a simple length penalty. Error ribbons are bootstrapped 95% CIs.

interactive speaking task, using the same sequence of images and feedback about correctness. We found that that the ablated model with no augmentation fails to become more efficient: at least for our RNN decoder architecture, evidence of success only reinforces expectations about the full caption; it cannot not propagate this evidence to the individual parts of the utterance. A sufficiently high length penalty does allow utterances to become shorter, but reduces linearly rather than quadratically (as humans do; Clark and Wilkes-Gibbs, 1986; Hawkins et al., 2019). Moreover, upon inspecting the utterances produced by this variant (see Table S1), we found that it simply cuts off the ends of utterances, whereas compositional data augmentation is based on a syntactic parse and thus allows the model to preserve grammaticality and gradually build expectations about meaningful sub-units. In sum, we find that our augmentation mechanism is not reducible to a coarse length bias, and is able to compensate for representation failures in current recurrent architectures (Dasgupta et al., 2018; Nikolaus et al., 2019).[1]

References

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

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639.

Kanishk Gandhi and Brenden M Lake. 2019. Mutual exclusivity as a challenge for neural networks. *arXiv preprint arXiv:1906.10197*.

Robert D Hawkins, Michael C Frank, and Noah D Goodman. 2019. Characterizing the dynamics of learning in repeated reference games. *arXiv preprint arXiv:1912.07199*.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *arXiv:2001.03632 [cs]*. ArXiv: 2001.03632.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. *arXiv preprint arXiv:1909.04402*.

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

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

[1] We expect that a more structured architecture, able to propagate evidence about the meaning of a full utterance to representations of intermediate semantic units, would make this augmentation step redundant (e.g. Tai et al., 2015; Gan et al., 2017; McCoy et al., 2020).