Curiosity in self-supervised active word learning

Lieke Gelderloos (Tilburg University), Alireza Mahmoudi Kamelabad (University of Trento), Afra Alishahi (Tilburg University)
l.j.gelderloos@uvt.nl

Cognitive models of child word learning in general, and cross-situational models in particular, characterize the learner as a passive observer. But children are curious and actively participate in verbal and non-verbal communication, and often introduce new topics which parents are likely to follow up (Bloom et al., 1996). We investigate the potential impact of curiosity on word learning through a series of computational experiments. Our simulation results show that a curious learner who actively and curiously influences the language input it receives learns faster and more robustly, and reaches better performance.

Model

We present a computational model that learns to map words to objects through language production and comprehension; see Figure 3 for a sketch of the architecture. The comprehension module decides which object in a given scene a word refers to, whereas the production module outputs a word for a given object. Scenes, objects and words come from real images (we used the dataset proposed by Keijser et al., 2019), and objects are represented as continuous, high level visual vectors (Simonyan and Zisserman, 2015).

The output of each module can serve as input to the other: when the production module outputs a word the comprehension module tries to interpret it, and when the comprehension module selects an object as referent for a word the production module tries to name the object. This property is used to train the model through self-supervision.

In our computational experiments, we compare the performance of a 'passive' model which receives the label of a random object of a given scene, with an 'active' model which chooses to receive the label for a target object with the highest learning potential. The metrics we use for estimating learning potential are subjective novelty, plasticity, and curiosity. Subjective novelty favours the most unknown objects, plasticity selects on how much the learner expects to learn from a given input word, and curiosity is the product of those two.

Results

Table 1 shows model accuracy in each condition on held-out test data. Models trained with input selection according to curiosity ultimately attain the highest accuracy scores for both comprehension and production. However, neither models trained with plasticity nor subjective novelty as selection mechanism outperform models trained with random object selection. Notably, the standard deviation is also smallest for models in the curious condition.

Figures 1 and 2 show the intermediate scores during training of all models on training and test data. Figure 2 shows that all models are prone to overfitting on the production task. Nevertheless, we also see clear differences in test scores between the conditions. For both comprehension and production, models in the curious condition immediately outperform those in other conditions, with models in the random condition gradually catching up. Models trained in the subjective novelty condition are prone to overfitting on both tasks.

Our results suggest that active participation of a word learner in input selection can have a noticeable impact on their accuracy, stability and learning trajectories. Whether learners do employ such a strategy must be established in empirical research.

Table 1: Average accuracy on test data

	Comprehension		Production	
	Acc.	SD	Acc.	SD
Random	.5458	.0746	.2093	.0139
Plasticity	.5119	.1035	.1801	.0223
Subjective novelty	.2874	.0026	.1214	.0082
Curiosity	.6626	.0190	.2132	.0046
Baseline	.2863		.0893	

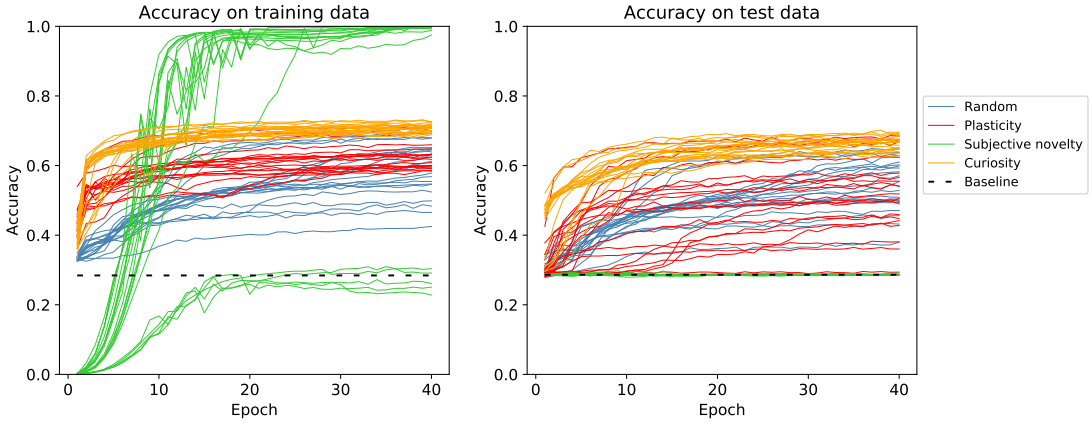

Figure 1: Comprehension train and test accuracy during training.

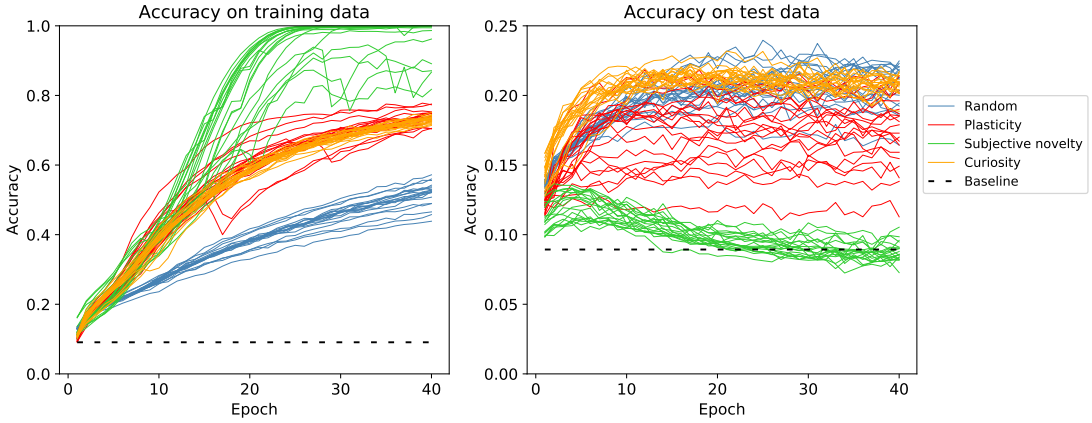

Figure 2: Production train and test accuracy during training. Please note that the y-axis for test accuracy is scaled for visibility.

References

Bloom, L., Margulis, C., Tinker, E., and Fujita, N. (1996). Early conversations and word learning: Contributions from child and adult. *Child Development*, 67(6):3154–3175.

Keijser, D., Gelderloos, L., and Alishahi, A. (2019). Curious topics: A curiosity-based model of first language word learning. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 1991–1997. Cognitive Science Society.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*.

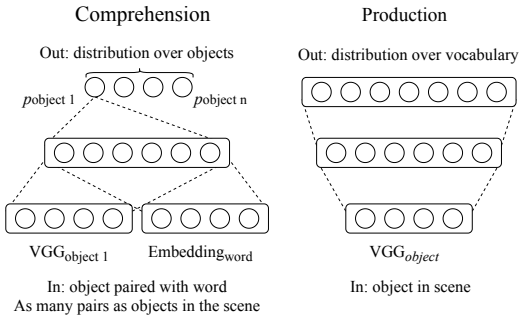

Figure 3: Architecture of the model, adapted from Keijser et al. (2019)

.

Technical details

Comprehension module. The candidacy of every object in a scene as the referent of a given word is considered in parallel: when the module receives a word as input, it concatenates the word embedding to the VGG vector of every object in the scene separately. This concatenated vector is input to a 256-unit hidden layer followed by a sigmoid activation function, which is fully connected to a single output unit also followed by sigmoid activation, and the object with the highest output value is selected as referent.

Production module. The input is a VGG vector, fed to a 256 unit hidden layer followed by sigmoid activation which is fully connected to the vocabulary-sized output layer (where every unit represents a word). The word unit with the highest output value is the best candidate to describe the target object.

Training. Once a word is processed by the comprehension module, we use the softmaxed output vector as attention over the objects in the scene. Input to the production module consists of the sum of the visual feature vectors of all the objects in the scene, weighted by the output of the comprehension module. The whole agent is updated in one go, according to the cross-entropy between the one-hot encoding of the input word and the output of the production module.

Definitions of curiosity. All metrics of curiosity are calculated through introspection: for each object in the scene, the learner uses the production module to produce a label, and then interprets the label using the comprehension module. Subjective novelty is defined as

$$s(t,o) = \frac{\sum_{i=1}^{n} \left(|t_i - o_i|\right))}{n} \tag{1}$$

where t is a one-hot vector encoding the object we are calculating subjective novelty for, o is the guessed output by the comprehension module, and n is the number of objects in the scene. Since every element of o is the result of a sigmoid function, $o_i\,(1-o_i)$ is the derivative of o_i. Because model updates are based on this value, plasticity uses it to estimate learning potential:

$$p(o) = \frac{\sum_{i=1}^{n} o_i\,(1-o_i)}{n} \tag{2}$$

Curiosity is the product of subjective novelty and plasticity, averaged over objects in the scene:

$$c(t,o) = \frac{\sum_{i=1}^{n} \left(|t_i - o_i|\right) o_i\,(1-o_i)}{n} \tag{3}$$