Like a Baby: Visually Situated Neural Language Acquisition

Alexander G. Ororbia*[1,2], **Ankur Mali** *[1,2], **Matthew A. Kelly**[1], and **David Reitter**[1,3]
(1) The Pennsylvania State University, University Park, PA, USA
(2) Rochester Institute of Technology, Rochester, NY, USA
(3) Google Research, New York City, NY, USA
ago@cs.rit.edu, aam35@psu.edu,
matthew.kelly@psu.edu, reitter@google.com

Abstract

We examine the benefits of visual context in training neural language models to perform next-word prediction. A multi-modal neural architecture is introduced that outperform its equivalent trained on language alone with a 2% decrease in perplexity, even when no visual context is available at test. Fine-tuning the embeddings of a pre-trained state-of-the-art bidirectional language model (BERT) in the language modeling framework yields a 3.5% improvement. The advantage for training with visual context when testing without is robust across different languages (English, German and Spanish) and different models (GRU, LSTM, Δ-RNN, as well as those that use BERT embeddings). Thus, language models perform better when they learn like a baby, i.e, in a multi-modal environment. This finding is compatible with the theory of situated cognition: language is inseparable from its physical context.

1 Introduction

The theory of situated cognition postulates that a person's knowledge is inseparable from the physical or social context in which it is learned and used (Greeno and Moore, 1993). Similarly, Perceptual Symbol Systems theory holds that all of cognition, thought, language, reasoning, and memory, is grounded in perceptual features (Barsalou, 1999). Knowledge of language cannot be separated from its physical context, which allows words and sentences to be learned by grounding them in reference to objects or natural concepts on hand (see Roy and Reiter, 2005, for a review). Nor can knowledge of language be separated from its social context, where language is learned interactively through communicating with others to facilitate problem-solving. Simply put, language does not occur in a vacuum.

Yet, statistical language models, typically connectionist systems, are often trained in such a vacuum. Sequences of symbols, such as sentences or phrases composed of words in any language, such as English or German, are often fed into the model independently of any real-world context they might describe. In the classical language modeling framework, a model learns to predict a word based on a history of words it has seen so far. While these models learn a great deal of linguistic structure from these symbol sequences alone, acquiring the essence of basic syntax, it is highly unlikely that this approach can create models that acquire much in terms of semantics or pragmatics, which are integral to the human experience of language. How might one build neural language models that "understand" the semantic content held within the symbol sequences, of any language, presented to it?

In this paper, we take a small step towards a model that understands language as a human does by training a neural model jointly on corresponding linguistic and visual data. From an image-captioning dataset, we create a multi-lingual corpus where sentences are mapped to the real-world images they describe. We ask how adding such real-world context at training can improve language model performance. We create a unified multi-modal connectionist architecture that incorporates visual context and uses either Δ-RNN (Ororbia II et al., 2017), Long Short Term Memory (LSTM; Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (GRU; Cho et al., 2014) units. We find that the models acquire more knowledge of language than if they were trained without corresponding, real-world visual context.

2 Related Work

Both behavioral and neuroimaging studies have found considerable evidence for the contribution of perceptual information to linguistic tasks (Barsalou, 2008). It has long been held that language is acquired jointly with perception through interaction with the environment (e.g. Frank et al., 2008). Eye-tracking studies show that visual context influences word recognition and syntactic parsing from even the earliest moments of comprehension (Tanenhaus et al., 1995).

Computational cognitive models can account for bootstrapped learning of word meaning and syntax when language is paired with perceptual experience (Abend et al., 2017) and for the ability of children to rapidly acquire new words by inferring the referent from their physical environment (Alishahi et al., 2008). Some distributional semantics models integrate word co-occurrence data with perceptual data, either to achieve a better model of language as it exists in the minds of humans (Baroni, 2016; Johns and Jones, 2012; Kievit-Kylar and Jones, 2011; Lazaridou et al., 2014) or to improve performance on machine learning tasks such as object recognition (Frome et al., 2013; Lazaridou et al., 2015a), image captioning (Kiros et al., 2014; Lazaridou et al., 2015b), or image search (Socher et al., 2014).

Integrating language and perception can facilitate language acquisition by allowing models to infer how a new word is used from the perceptual features of its referent (Johns and Jones, 2012) or to allow for *fast mapping* between a new word and a new object in the environment (Lazaridou et al., 2014). Likewise, this integration allows models to infer the perceptual features of an unobserved referent from how a word is used in language (Johns and Jones, 2012; Lazaridou et al., 2015b). As a result, language data can be used to improve object recognition by providing information about unobserved or infrequently observed objects (Frome et al., 2013) or for differentiating objects that often co-occur in photos (e.g., cats and sofas; Lazaridou et al., 2015a).

By representing the referents of concrete nouns as arrangements of elementary visual features (Biederman, 1987), Kievit-Kylar and Jones (2011) found that the visual features of nouns capture semantic typicality effects, and that a combined representation, consisting of both visual features and word co-occurrence data, more strongly correlates with human judgments of semantic similarity than representations extracted from a corpus alone. While modeling similarity judgments is distinct from the problem of predictive language modeling, we take this finding as evidence that visual perception informs semantics, which suggests there are gains to be had integrating perception with predictive language models.

In contrast to prior work in machine learning, where mappings between vision and language have been examined (Kiros et al., 2014; Vinyals et al., 2015; Xu et al., 2015), our goal in integrating visual and linguistic data is not to accomplish a task such as image search/captioning that inherently requires a mapping between these modalities. Rather, our goal is to show that, since perceptual information is intrinsic to how humans process language, a language model that is trained on both visual and linguistic data will be a better model, consistently across languages, than one trained on linguistic data alone.

Due to the ability of language models to constrain predictions on the basis of preceding context, language models play a central role in natural-language and speech processing applications. However, the psycholinguistic questions surrounding how people acquire and use linguistic knowledge are fundamentally different from the aims of machine learning. Using NLP-style language models to address psycholinguistic questions is a new approach that integrates well with the theory of predictive coding in cognitive psychology (Clark, 2013; Rao and Ballard, 1999). For language processing this means that when reading text or comprehending speech, humans constantly anticipate what will be said next. Predictive coding in humans is a fast, implicit cognitive process similar to the kind of sequence learning that recurrent neural models excel at. We do not propose recurrent neural models as direct accounts of human language processing. Instead, our intent is to use a general purpose machine learning algorithm as a tool to investigate the informational characteristics of the language learning task. More specifically, we use machine learning to explore the question as to whether natural languages are most easily learned when situated in an environmental context and grounded in perception.

3 The Multi-modal Neural Architecture

We will evaluate the multi-modal training approach on several well-known complex architectures, including the LSTM, and further examine the effect of using pre-trained BERT embeddings. However, to simply describe the the neural model, we start from the Differential State Framework (DSF; Ororbia II et al., 2017), which unifies gated recurrent architectures under the general view that state memory is a simple parametrized mixture of "fast" and "slow" states. Our aim is to model sequences of symbols, such as the words that compose sentences, where at each time we process $\mathbf{x}_t$, or the one-hot encoding of a token[1]

One of the simplest models that can be derived from the DSF is the Δ-RNN (Ororbia II et al., 2017). A Δ-RNN is a simple gated RNN that captures longer-term dependencies in sequences through the use of a parametrized, flexible state "mixing" function. The model computes a new state at a given time step by comparing a fast state (which is proposed after accounting for the current token) and a slow state (a form of long-term memory). The model is defined by parameters $\Theta = \{W, V, \mathbf{b}_r, \beta_1, \beta_2, \alpha\}$ (input-to-hidden weights W, recurrent weights V, gating-control coefficients β_1, β_2, α, and the rate-gate bias $\mathbf{b}_r$). Inference is defined as:

$$\mathbf{d}_t^{rec} = V\mathbf{h}_{t-1},\ \mathbf{d}_t^{dat} = W\mathbf{e}_{w,t} \quad (1)$$

$$\mathbf{d}_t^1 = \alpha \otimes \mathbf{d}_t^{rec} \otimes \mathbf{d}_t^{dat} \quad (2)$$

$$\mathbf{d}_t^2 = \beta_1 \otimes \mathbf{d}_t^{rec} + \beta_2 \otimes \mathbf{d}_t^{dat} \quad (3)$$

$$\mathbf{z}_t = \phi_{hid}(\mathbf{d}_t^1 + \mathbf{d}_t^2) \quad (4)$$

$$\mathbf{h}_t = \Phi((1 - \mathbf{r}) \otimes \mathbf{z}_t + \mathbf{r} \otimes \mathbf{h}_{t-1}) \quad (5)$$

$$\mathbf{r} = 1/(1 + exp(-[\mathbf{d}_t^{dat} + \mathbf{b}_r])) \quad (6)$$

where $\mathbf{e}_{w,t}$ is the 1-of-k encoding of the word w at time t. Note that $\{\alpha, \beta_1, \beta_2\}$ are learnable bias vectors that modulate internal multiplicative interactions. The rate gate $\mathbf{r}$ controls how slow and fast-moving memory states are mixed inside the model. In contrast to the model originally trained in Ororbia II et al. (2017), the outer activation is the linear rectifier, $\Phi(v) = max(0, v)$, instead of the identity or hyperbolic tangent, because we found that it worked much better. The inner activation function $\phi_{hid}(v)$ is $tanh(v) = \frac{(e^{(2v)}-1)}{(e^{(2v)}+1)}$.

To integrate visual context information into the Δ-RNN, we fuse the model with a neural vision system, motivated by work done in automated image captioning (Xu et al., 2015). We adopt a transfer learning approach and incorporate a state-of-the-art convolutional neural network into the Δ-RNN model, namely the Inception-v3 network (Szegedy et al., 2016)[2], in order to create a multi-modal Δ-RNN model (MM-Δ-RNN; see Figure 1). Since our focus is on language modeling, the parameters of the vision network are fixed.

To obtain a distributed representation of an image from the Inception-v3 network, we extract the vector produced from the final max-pooling layer, $\mathbf{c}$, after running an image through the model (note that this operation occurs right before the final, fully-connected processing layers which are usually task-specific parameters, such as in object classification). The Δ-RNN can make use of the information in this visual context vector if we modify its state computation in one of two ways. The first way would be to modify the inner state to be a linear combination of the data-dependent pre-activation, the filtration, and a learned linear mapping of $\mathbf{c}$ as follows:

$$\mathbf{z}_t = \phi_{hid}(\mathbf{d}_t^1 + \mathbf{d}_t^2 + M\mathbf{c} + \mathbf{b}) \quad (7)$$

where M is a learnable synaptic connections matrix that connects the visual context representation with the inner state. The second way to modify the Δ-RNN would be change its outer mixing function instead:

$$\mathbf{h}_t = \Phi([(1 - \mathbf{r}) \otimes \mathbf{z}_t + \mathbf{r} \otimes \mathbf{h}_{t-1}] \otimes (M\mathbf{c})) \quad (8)$$

Here in Equation 8 we see the linearly-mapped visual context embedding interacts with the currently computation state through a multiplicative operation, allowing the visual-context to persist and work in a longer-term capacity. In either situation, using a parameter matrix M frees us from having to set the dimensionality of the hidden state to be the same as the context vector produced by the Inception-v3 network.

We do not use regularization techniques with this model. The application of regularization techniques is, in principle, possible (and typically im-

[1] One-hot encoding represents tokens as binary-valued vectors with one dimension for each type of token. Only one dimension has a non-zero value, indicating the presence of a token of that type.

[2] In preliminary experiments, we also examined VGGNet and a few others, but found that the Inception worked the best when it came to acquiring more general distributed representations of natural images.

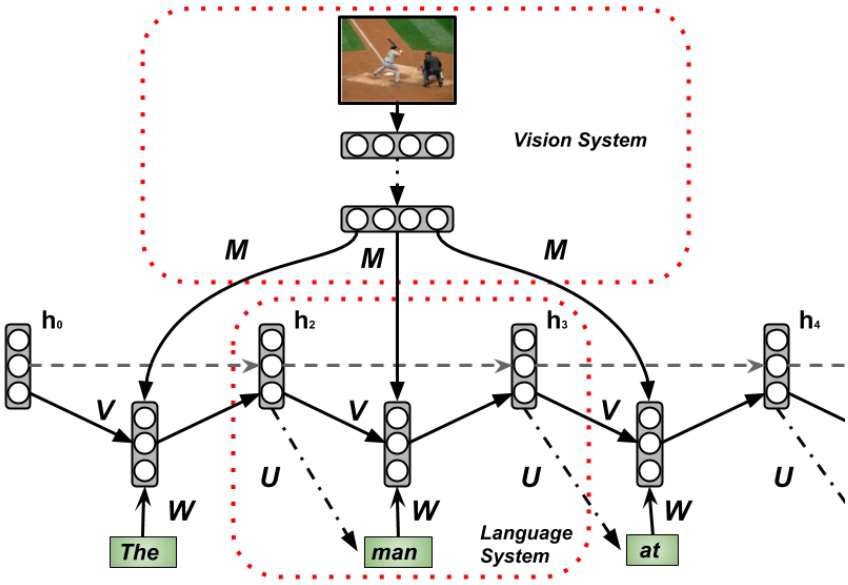


Figure 1: Integration of visual information in an unrolled network (here, the MM-Δ-RNN. Grey-dashed: identity connections; black-dash-dotted: next-step predictions; solid-back lines: weight matrices.

proves performance of the Δ-RNN), but it is damaging to performance in this particular case, where an already compressed and regularized representation of the images from Inception-v3 serves as input to the multi-modal language modeling network.

Let $w_1, \ldots, w_N$ be a variable-length sequence of N words corresponding to an image I. In general, the distribution over the variables follows the graphical model:

$$P_\theta(w_1, \ldots, w_T | I) = \prod_{t=1}^{T} P_\Theta(w_t | w_{<t}, I)$$

For all model variants the state $\mathbf{h}_t$ calculated at any time step is fed into a maximum-entropy classifier[3] defined as:

$$P(w, \mathbf{h}_t) = P_\Theta(w|\mathbf{h}_t) = \frac{\exp\left(w^{\mathrm{T}} U \mathbf{h}_t\right)}{\sum_{w'} \exp\left((w')^{\mathrm{T}} U \mathbf{h}_t\right)}$$

The model parameters Θ optimized with respect to the sequence negative log likelihood:

$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{t=1}^{T} \log P_\Theta(w_t | \mathbf{h})$$

We differentiate with respect to this cost function to calculate gradients.

[3]Bias term omitted for clarity.

3.1 GRU, LSTM and BERT variants

Does visually situated language learning benefit from the specific architecture of the Δ-RNN, or does the proposal work with state-of-the-art language models? We applied the same architecture to Gated Recurrent Units (GRU, Cho et al., 2014), Long Short Term Memory (LSTM, Hochreiter and Schmidhuber, 1997), and BERT (Devlin et al., 2018). We train these models on text alone and compare to the two variations of the multi-modal Δ-RNN, as described in the previous section. The multi-modal GRU, with context information directly integrated, is defined as follows:

$$\begin{aligned}
\mathbf{d}_c &= M\mathbf{c} \\
\mathbf{z}_t &= \sigma(W_z \mathbf{x}_t + V_z \mathbf{h}_{t-1}) \\
\mathbf{r}_t &= \sigma(W_r \mathbf{x}_t + V_r \mathbf{h}_{t-1}) \\
\widehat{\mathbf{h}}_t &= \tanh(W_{\widehat{h}} \mathbf{x}_t + V_{\widehat{h}}(\mathbf{r}_t \otimes \mathbf{h}_{t-1})) \\
\mathbf{h}_t &= [\mathbf{z}_t \otimes \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \otimes \widehat{\mathbf{h}}_t] \otimes \mathbf{d}_c
\end{aligned}$$

where we note the parameter matrix M that maps the visual context c into the GRU state effectively gates the outer function.[4] The multi-modal variant of the LSTM (with peephole connections) is

[4]We tried both methods of integration, Equations 7 and 8. The second formulation gave better performance.

defined as follows:

$$\begin{aligned}
\mathbf{d}_c &= M\mathbf{c} \\
\mathbf{h}_t &= [\mathbf{r}_t \otimes \Phi(\mathbf{c}_t)] \otimes \mathbf{d}_c, \text{ where,} \\
\mathbf{r}_t &= \sigma(W_r\mathbf{x}_t + V_r\mathbf{h}_{t-1} + U_r\mathbf{c}_t) \\
\mathbf{c}_t &= \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \mathbf{z}_t, \text{ where,} \\
\mathbf{z}_t &= \Phi(W_z\mathbf{x}_t + V_z\mathbf{h}_{t-1}), \\
\mathbf{i}_t &= \sigma(W_i\mathbf{x}_t + V_i\mathbf{h}_{t-1} + U_i\mathbf{c}_{t-1}), \\
\mathbf{f}_t &= \sigma(W_f\mathbf{x}_t + V_f\mathbf{h}_{t-1} + U_f\mathbf{c}_{t-1}).
\end{aligned}$$

We furthermore created one more variant of each multi-modal RNN by initializing a portion of their input-to-hidden weights with embeddings extracted from the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018). This would correspond to initializing W in the Δ-RNN, W_i in the LSTM, and $W_{\hat{h}}$ in the GRU. Note that in our results, we only report the best-performing model, which turned out to be the LSTM variant. Since the models in this work are at the word level and BERT operates at the subword level, we create initial word embeddings by first decomposing each word into its appropriate subword components, according to the WordPieces model (Wu et al., 2016), and then extract the relevant BERT representation for each. For each subword token, a representation is created by summing together a specific learned token embedding, a segmentation embedding, and a position embedding. For a target word, we linearly combine subword input representations and initialize the relevant weight with this final embedding.

4 Experiments

The experiments in this paper were conducted using the MS-COCO image-captioning dataset.[5] Each image in the dataset has five captions provided by human annotators. We use the captions to create five different ground truth splits. We translated each ground truth split into German and Spanish using the Google Translation API, which was chosen as a state-of-the-art, independently evaluated MT tool that produces, according to our inspection of the results, idiomatic, and syntactically and semantically faithful translations. To our knowledge, this represents the first Multi-lingual MSCOCO dataset on situated learning. We tokenize the corpus and obtain a 16.6K vocabulary for English, 33.2K for German and 18.2k for Spanish.

[5] https://competitions.codalab.org/competitions/3221

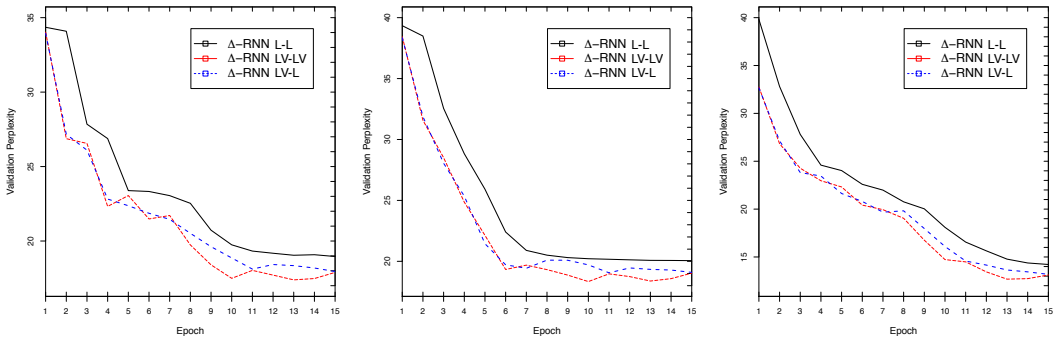
As our primary concern is the next-step prediction of words/tokens, we use negative log likelihood and perplexity to evaluate the models. This is different from the goals of machine translation or image captioning, which, in most cases, is concerned with a ranking of possible captions where one measures how similar the model's generated sequences are to ground-truth target phrases.

Baseline results were obtained with neural language models trained on text alone. For the Δ-RNN, this meant implementing a model using only Equations 1-7. The best results were achieved using the BERT Large model (bidirectional Transformer, 24 layers, 1024dims, 16 attention heads: Devlin et al. 2018). We used the large pretrained model and then trained with visual context.

All models were trained to minimize the sequence loss of the sentences in the training split. The weight matrices of all models were initialized from uniform distribution, $U(-0.1, 0.1)$, biases were initialized from zero, and the Δ-RNN-specific biases $\{\alpha, \beta_1, \beta_2\}$ were all initialized to one. Parameter updates calculated through backpropagation through time required unrolling the model over 49 steps in time (this length was determined based on validation set likelihood). All symbol sequences were zero-padded and appropriately masked to ensure efficient mini-batching. Gradients were hard-clipped at a magnitude bound of $l = 2.0$. Over mini-batches of 32 samples, model parameters were optimized using simple stochastic gradient descent (learning rate $\lambda = 1.0$ which was halved if the perplexity, measured at the end of each epoch, goes up three or more times).

To determine if our multi-modal language models capture knowledge that is different from a text-only language model, we evaluate each model twice. First, we compute the model perplexity on the test set using the sentences' visual context vectors. Next, we compute model perplexity on test sentences by feeding in a null-vector to the multi-modal model as the visual context. If the model did truly pick up some semantic knowledge that is not exclusively dependent on the context vector, its perplexity in the second setting, while naturally worse than the first setting, should still outperform text-only baselines.

In Table 1, we report each model's negative log likelihood (NLL) and per-word perplexity (PPL).



(a) English Δ-RNNs. (b) German Δ-RNNs. (c) Spanish Δ-RNNs.

Figure 2: Training Δ-RNNs in each language (English, German, Spanish). Baseline model is trained and evaluated on language (L-L), the *full* model uses the multi-modal signal (LV-LV), and the target model is trained on LV, but evaluated on L only (LV-L).

PPL is calculated as:

$$PPL = \exp\Big[-(1/N)\sum_{i=1}^{N}\sum_{t=1}^{T}\log P_\Theta(w_t|\mathbf{h})\Big]$$

We observe that in all cases the multi-modal models outperform their respective text-only baselines. More importantly, the multi-modal models, when evaluated without the Inception-v3 representations on holdout samples, still perform better than the text-only baselines. The improvement in language generalization can be attributed to the visual context information provided during training, enriching its representations over word sequences with knowledge of actual objects and actions.

Figure 2 shows the validation perplexity of the Δ-RNN on each language as a function of the first 15 epochs of learning. We observe that throughout learning, the improvement in generalization afforded by the visual context $\mathbf{c}$ is persistent. Validation performance was also tracked for the various GRU and LSTM models, where the same trend was also observed.

4.1 Model Analysis

We analyze the decoders of text-only and multi-modal models. We examine the parameter matrix U, which is directly involved in calculating the predictions of the underlying generative model. U can be thought of as "transposed embeddings", an idea that has also been exploited to introduce further regularization into the neural language model learning process (Press and Wolf, 2016; Inan et al., 2016). If we treat each row of this matrix as the learned embedding for a particular word (we assume column-major orientation in implementation), we can calculate its proximity to other embeddings using cosine similarity.

Table 2 shows the top ten words for several randomly selected query terms using the decoder parameter matrix. By observing the different sets of nearest-neighbors produced by the Δ-RNN and the multi-modal Δ-RNN (MM-Δ-RNN), we can see that the MM-Δ-RNN appears to have learned to combine the information from the visual context with the token sequence in its representations. For example, for the query "ocean", we see that while the Δ-RNN does associate some relevant terms, such as "surfing" and "beach", it also associates terms with marginal relevance to "ocean" such as "market" and "plays". Conversely, nearly all of the terms the MM-Δ-RNN associates with "ocean" are relevant to the query. The same is true for "kite" and "subway". For "racket", while the text-only baseline mostly associates the query with sports terms, especially sports equipment like "bat", the MM-Δ-RNN is able to relate the query to the correct sport, "tennis".

4.2 Conditional Sampling

To see how visual context influences the language model, we sample the conditional generative model. Beam search (size 13) allows us to generate full sentences (Table 3). Words were ranked based on model probabilities.

Table 1: Generalization performance as measured by negative log likelihood (NLL) and perplexity (PPL). Lower values indicate better performance. Baseline model (L-L) trained and evaluated on linguistic data only. Full model (LV-LV) trained and evaluated on both linguistic and visual data. Blind model (LV-L) trained on both but evaluated on language only. **The difference between L-L and LV-L illustrates the performance improvement.** German and Spanish data are machine-translated (MT) and provide additional, but correlated, evidence. For comparison, Devlin et al. (2018) report a perplexity of 3.23 for their (broad) English test data, using the same base model we use here to define input representations.

	English		**German MT**		**Spanish MT**	
Model (Type)	Test-NLL	Test-PPL	Test-NLL	Test-PPL	Test-NLL	Test-PPL
Δ-RNN (L-L)	2.714	15.086	2.836	17.052	2.546	12.755
MM-Δ-RNN (LV-LV)	2.645	14.086	2.777	16.082	2.405	11.082
MM-Δ-RNN (LV-L)	2.694	14.786	2.808	16.582	2.458	11.682
GRU (L-L)	2.764	15.871	2.854	17.369	2.554	12.866
MM-GRU (LV-LV)	2.654	14.189	2.790	16.285	2.426	11.3089
MM-GRU (LV-L)	2.687	14.689	2.815	16.701	2.466	11.781
LSTM (L-L)	2.722	15.217	2.814	17.070	2.494	12.114
MM-LSTM (LV-LV)	2.645	14.089	2.773	16.001	2.405	11.081
MM-LSTM (LV-L)	2.708	15.002	2.822	16.806	2.487	12.028
BERT+LSTM (L-L)	2.534	12.6011	2.702	14.9127	2.303	10.0011
BERT+MM-LSTM (LV-LV)	2.475	11.8776	2.661	14.3124	2.223	9.2319
BERT+MM-LSTM (LV-L)	**2.503**	**12.2196**	**2.700**	**14.8102**	**2.283**	**9.8102**

5 Discussion and Conclusions

Perceptual context improves training multi-modal neural models compared to training on language alone. Specifically, augmenting a predictive language model with images that illustrate the sentences being learned enhances its next-word prediction ability. The performance improvement persists even in situations devoid of visual input, when the model is used as a pure language model.

The near state-of-the-art language model, using BERT, reflects the case of human language acquisition less than do the other models, which were trained "ab initio" in a situated context. BERT is pre-trained on a very large corpus, but it still picked up a performance improvement when fine-tuned on the visual context and language, as compared to the corpus language signal alone. We do not expect this to be a ceiling for visual augmentation: in the world of training LMs, the MS COCO corpus is, of course, a small dataset.

Neural language models, as used here, are contenders as cognitive and psycholinguistic models of the non-symbolic, implicit aspects of language representation. There is a great deal of evidence that something like a predictive language model exists in the human mind. The *surprisal* of a word or phrase refers to the degree of mismatch between what a human listener expected to be said next and what is actually said, for example, when a garden path sentence forces the listener to abandon a partial, incremental parse (Ferreira and Henderson, 1991; Hale, 2001). In the garden path sentence "The horse raced past the barn fell", the final word "fell" forces the reader to revise their initial interpretation of "raced" as the active verb (Bever, 1970). More generally, the idea of *predictive coding* holds that the mind forms expectations before perception occurs (see Clark, 2013, for a review). How these predictions are formed is unclear. Predictive language models trained with a generic neural architecture, without specific linguistic universals, are a reasonable candidate for a model of predictive coding in language. This does not imply neuropsychological realism of the low-level representations or learning algorithms, and we cannot advocate for a specific neural architecture as being most plausible. However, we can show that an architecture that predicts linguistic input well learns better when its input mimics that of a human language learner.

Table 2: The ten words most closely related to the bolded query word, rank ordered, trained with (MM-Δ-RNN) and without (Δ-RNN) visual input.

Ocean		**Kite**		**Subway**		**Racket**	
Δ-RNN	+MM	**Δ-RNN**	+MM	**Δ-RNN**	+MM	**Δ-RNN**	+MM
surfing	boats	plane	kites	train	railroad	bat	bat
sandy	beach	kites	airplane	passenger	train	batter	players
filled	pier	airplane	plane	railroad	locomotive	catcher	batter
beach	wetsuit	surfboard	airplanes	trains	trains	skateboard	swing
market	cloth	planes	planes	gas	steam	umpire	catcher
crowded	surfing	airplanes	airliner	commuter	gas	soccer	hitter
topped	windsurfing	boats	helicopter	trolley	commuter	women	ball
plays	boardwalk	jet	jets	locomotive	passenger	pedestrians	umpire
cross	flying	aircraft	biplane	steam	crowded	players	tennis
snowy	biplane	jets	jet	it's	trolley	uniform	tatoos

A cognitive model of language processing might distinguish between symbolic language knowledge and processes that implement compositionality to produce semantics on the one hand, and implicit processes that leverage sequences and associations to produce expectations. With respect to acquiring the latter, implicit and predictive model, we note that children are exposed to a rich sensory environment, one more detailed than the environment provided to our model here. If even static visual input alone improves language acquisition, then what could a sensorily rich environment achieve? When a multi-modal learner is considered, then, perhaps, the language acquisition stimulus that has been famously labeled to be rather poor (Chomsky, 1959; Berwick et al., 2013), is not so poor after all.

Table 3: Some captions generated by the multi-modal Δ-RNN in English.

a skateboarder and person in front of skyscrapers.
a person with skateboarder on air.
a person doing a trick with skateboarder.
a person with camera with blue background.

a food bowl on the table
a bowl full of food on the table
a green and red bowl on the table
a salad bowl with chicken

a dog on blue bed with blanket.
a dog sleeps near wooden table.
a dog sleeps on a bed.
a dog on some blue blankets.

Acknowledgments

We would like to thank Tomas Mikolov, Emily Pitler, Zixin Tang, and Saranya Venkatraman for comments. Part of this work was funded by the National Science Foundation (BCS-1734304 to D. Reitter).

References

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116 – 143.

Afra Alishahi, Afsaneh Fazly, and Suzanne Stevenson. 2008. Fast mapping in word learning: What probabilities tell us. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 57–64. Association for Computational Linguistics.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Lawrence W Barsalou. 1999. Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, 22(4):637–660.

Lawrence W Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Robert C Berwick, Noam Chomsky, and Massimo Piattelli-Palmarini. 2013. Poverty of the stimulus

stands: Why recent challenges fail. In *Rich Languages From Poor Inputs*, chapter 1, pages 19–42. Oxford University Press.

Thomas G Bever. 1970. The cognitive basis for linguistic structures. In *Cognition and the development of language*, pages 279–362.

Irving Biederman. 1987. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Noam Chomsky. 1959. A review of BF Skinner's verbal behavior. *Language*, 35(1):26–58.

Andy Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fernanda Ferreira and John M Henderson. 1991. Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6):725–745.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2008. A Bayesian framework for cross-situational word-learning. In *Advances in neural information processing systems*, pages 457–464.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129.

James G Greeno and Joyce L Moore. 1993. Situativity and symbols: Response to Vera and Simon. *Cognitive Science*, 17(1):49–59.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8, Pittsburgh, PA.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2016. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462*.

Brendan T Johns and Michael N Jones. 2012. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1):103–120.

Brent Kievit-Kylar and Michael Jones. 2011. The semantic pictionary project. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 2229–2234, Austin, TX. Cognitive Science Society.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1403–1414.

Angeliki Lazaridou, Georgiana Dinu, Adam Liska, and Marco Baroni. 2015a. From visual attributes to adjectives through decompositional distributional semantics. *Transactions of the Association for Computational Linguistics*, 3:183–196.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015b. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Alexander G. Ororbia II, Tomas Mikolov, and David Reitter. 2017. Learning simpler language models with the differential state framework. *Neural Computation*, 29(12):3327–3352.

Ofir Press and Lior Wolf. 2016. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*.

Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218.

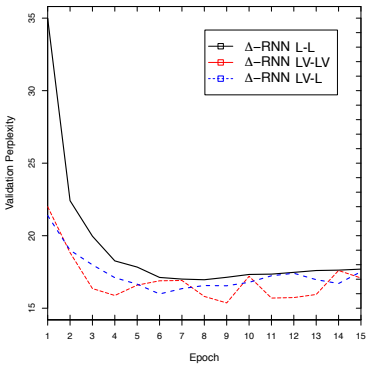
Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

MK Tanenhaus, MJ Spivey-Knowlton, KM Eberhard, and JC Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

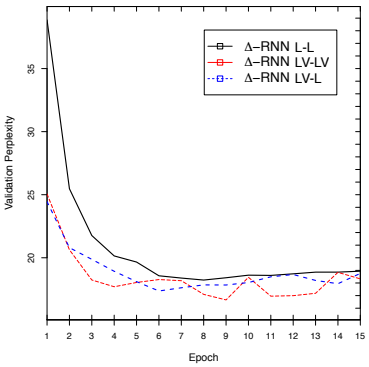
Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

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*.

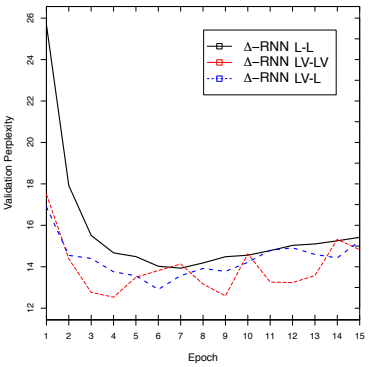
Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.



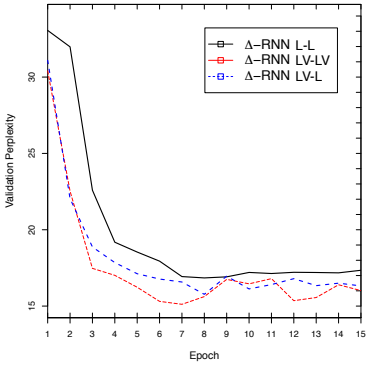
(a) English GRUs.



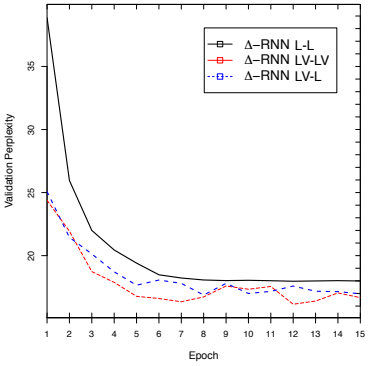
(b) German GRUs.



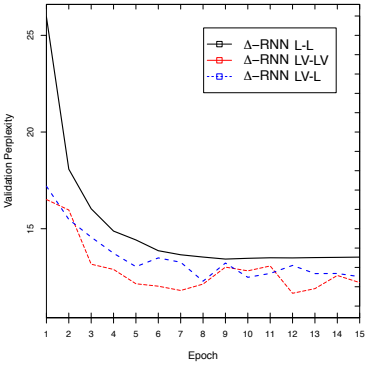
(c) Spanish GRUs.



(d) English LSTMs.



(e) German LSTMs.



(f) Spanish LSTMs.

Appendix: Comparison of learning curves for the GRUs and LSTMs in each language (English, German, Spanish). To augment Figure 2 in the main paper, we also show the learning curves for all models experimented with in this paper beyond the Δ-RNN. Validation learning curves are provided for the GRU and LSTM language models, both multimodal and unimodal variations.