Summary must stand on 10 lines max. please

Learning word meanings from images of natural scenes

Ákos Kádár * — **Afra Alishahi** * — **Grzegorz Chrupała** *

* *Tilburg Center for Cognition and Communication, Tilburg University*

S//////

unless you have footnote or other refere for these... bu I can't see them

ABSTRACT. *Children in their early life are faced with the challenge of learning the meanings of words from noisy and highly ambiguous contexts. The utterances that guide their learning are emitted in complex scenes where they need to identify which aspects of these scenes are related to which parts of the perceived utterances. One of the key challenges in computational modeling of the acquisition of word meanings is to provide rich enough representations of scenes that contain the similar sources of information and have similar statistical properties as naturally occurring data. In this paper we propose a novel computational model of cross-situational word learning that takes images of natural scenes paired with their descriptions as input and incrementally learns probabilistic associations between words and image features. We examine our model's ability to learn word meanings from ambiguous and noisy data through a set of experiments. Our results show that the model is able to learn meaning representations that correlate with human similarity judgments of word pairs. Furthermore, we show that given an image of a natural scene our model is able to name words conceptually related to the image.*

RÉSUMÉ. *Les jeunes enfants doivent apprendre le sens des mots à partir de contextes bruités et hautement ambigus. Les énoncés qu'ils reçoivent en phase d'acquisition sont produits lors de scènes complexes où ils doivent inférer quel mot réfère à quelle partie de la scène. Un des enjeux de la modélisation informatique de l'acquisition du sens des mots est donc de fournir des représentations très détaillées des scènes en question afin de contenir les mêmes sources de données et les mêmes propriétés statistiques que les données naturelles. Dans cet article, nous proposons un nouveau modèle informatique de l'apprentissage du sens des mots à partir d'images représentant des scènes liées à des descriptions détaillées. À partir de là, notre modèle apprend de manière incrémentale des associations statistiques entre mots et images : nous examinons ensuite, à travers une série d'expériences, dans quelle mesure ce modèle peut apprendre le sens des mots à partir de données bruitées. Nos résultats montrent que le modèle est capable d'apprendre des représentations sémantiques corrélées avec les jugements humains concernant la similarité de couples de mots. De plus, nous montrons qu'étant donné une scène précise, le modèle peut énumérer des mots conceptuellement liés à la scène en question.*

S/

permettre d'I

~~R^e soumission à TAL 55-3, le 15/06/2015~~ TAL. Volume 55 – n° 3/2014, pages 73 à 95

~~2 Re soumission à TAL 55-3~~ ⅃ ⊣

KEYWORDS: *Child language acquisition; Cross-situational learning; Computational Cognitive modeling; Multi-modal learning* ⅄ ·⅄

MOTS CLÉS : *Acquisition du langage par l'enfant ; apprentissage inter-situationnel ; modélisation informatique et sciences cognitives ; apprentissage multimodal* ⅄ ·⅄

These two pages must stand on only one, please.
You might have to shorten your summary...

Index of the signs used as proofreading devices:

⊣ ⅃ (or) /⅃ = suppress (or replace if there's a proposal next to "⊣" or "/"

⅄ = add

(bdc) = from capital (cap..) to small letter

] = attach to the line just before / or word

[= attach to the line / word after

I = stick before and after (usually suppress the space)

∽ = move from one place to another

(?) = suggestion ... you may not agree.

/J = from small "j" to capital "J"

L 74 TAL. Volume 55 – n°3/2014

1. Introduction

Children learn most of their vocabulary from hearing words in noisy and ambiguous contexts, where there are often many possible mappings between words and concepts. They attend to the visual environment to establish such mappings, but given that the visual context is often very rich and dynamic, elaborate cognitive processes are required for successful word learning from observation. Consider a language learner hearing the utterance *"the gull took my sandwich"* while watching a bird stealing someone's food. For the word *gull*, such information suggests potential mappings to the bird, the person, the action, or any other part of the observed scene. Further exposure to usages of this word and relying on structural cues from the sentence structure is necessary to narrow down the range of its possible meanings.

1.1. *Cross-situational learning*

A well-established account of word learning from perceptual context is called cross-situational learning, a bottom-up strategy in which the learner draws on the patterns of co-occurrence between a word and its referent across situations in order to reduce the number of possible mappings (Quine, 1960; Carey, 1978; Pinker, 1989). Various experimental studies have shown that both children and adults use cross-situational evidence for learning new words (Yu and Smith, 2007; Smith and Yu, 2008; Vouloumanos, 2008; Vouloumanos and Werker, 2009).

Cognitive word learning models have been extensively used to study how children learn robust word-meaning associations despite the high rate of noise and ambiguity in the input they receive. Most of the existing models are either simple associative networks that gradually learn to predict a word form based on a set of semantic features (Li *et al.*, 2004; Regier, 2005), or are rule-based or probabilistic implementations which use statistical regularities observed in the input to detect associations between linguistic labels and visual features or concepts (Siskind, 1996; Frank *et al.*, 2007; Yu, 2008; Fazly *et al.*, 2010). These models all implement different (implicit or explicit) variations of the cross-situational learning mechanism, and demonstrate its efficiency in learning robust mappings between words and meaning representations in presence of noise and perceptual ambiguity.

However, a main obstacle to developing realistic models of child word learning is lack of resources for reconstructing perceptual context. The input to a usage-based cognitive model must contain the same information components and statistical properties as naturally-occurring data children are exposed to. A large collection of transcriptions and video recordings of child-adult interactions has been accumulated over the years (MacWhinney, 2014), but few of these resources provide adequate semantic annotations that can be automatically used by a computational model. As a result, the existing models of word learning have relied on artificially generated input (Siskind, 1996). The meaning of each word is represented as a symbol or a set of semantic features that are selected arbitrarily or from lexical resources such as WordNet

You have to choose it yourself please...

(Fellbaum, 1998), and the visual context is built by sampling these symbols. Some models add additional noise to data by randomly adding or removing meaning symbols to/from the perceptual input (Fazly *et al.*, 2010).

Carefully constructed artificial input is useful in testing the plausibility of a learning mechanism, but comparisons with manually annotated visual scenes show that these artificially generated data sets often do not show the same level of complexity and ambiguity as naturally occurring perceptual context (Matusevych *et al.*, 2013; Beekhuizen *et al.*, 2013).

1.2. *Learning meanings from images*

To investigate the plausibility of cross-situational learning in a more naturalistic setting, we propose to use visual features from collections of images and their captions as input to a word learning model. In the domain of human-computer interaction (HCI) and robotics, a number of models have investigated the acquisition of terminology for visual concepts such as color and shape from visual data. Such concepts are learned based on communication with human users (Fleischman and Roy, 2005; Skocaj *et al.*, 2011). Because of the HCI setting, they need to make simplifying assumptions about the level of ambiguity and uncertainty about the visual context.

The input data we exploit in this research has been used for much recent work in NLP and machine learning whose goal is to develop multimodal systems for practical tasks such as automatic image captioning. This is a fast-growing field and a detailed discussion of it is beyond the scope of this paper. Recent systems include (Karpathy and Fei-Fei, 2014; Mao *et al.*, 2014; Kiros *et al.*, 2014; Donahue *et al.*, 2014; Vinyals *et al.*, 2014; Venugopalan *et al.*, 2014; Chen and Zitnick, 2014; Fang *et al.*, 2014). The majority of these approaches rely on convolutional neural networks for deriving representations of visual input, and then generate the captions using various versions of recurrent neural network language models conditioned on image representations. For example Vinyals *et al.* (2014) use the deep convolutional neural network of Szegedy *et al.* (2014) trained on ImageNet to encode the image into a vector. This representation is then decoded into a sentence using a Long Short-Term Memory recurrent neural network (Hochreiter and Schmidhuber, 1997). Words are represented by embedding them into a multidimensional space where similar words are close to each other. The parameters of this embedding are trainable together with the rest of the model, and are analogous to the vector representations learned by the model proposed in this paper. The authors show some example embeddings but do not analyze or evaluate them quantitatively, as their main focus is on the captioning performance.

Perhaps the approach most similar to ours is the model of Bruni *et al.* (2014). In their work, they train multimodal distributional semantics models on both textual information and bag-of-visual-words features extracted from captioned images. They use the induced semantic vectors for simulating word similarity judgments by humans, and show that a combination of text and image-based vectors can replicate human

2. Word learning model

Latest existing cross-situational models formulate word learning as a translation problem, where the learner must decide which words in an utterance correspond to which symbols (or potential referents) in the perceptual context (Yu and Ballard, 2007; Fazly *et al.*, 2010). For each new utterance paired with a symbolic representation of the visual scene, first the model decides which word is *aligned* with which symbol based on previous associations between the two. Next, it uses the estimated alignments to update the meaning representation associated with each word.

We introduce a novel computational model for cross-situational word learning from captioned images. We reformulate the problem of learning the meaning of words as a translation problem between words and a *continuous* representation of the scene; that is, the visual features extracted from the image. In this setting, the model learns word representations by taking images and their descriptions one pair at a time. To learn correspondences between English words and image features, we borrow and adapt the translation-table estimation component of the IBM Model 1 (Brown *et al.*, 1993). The learning results in a translation table between words and image-features, i.e. a list of probabilities of image-features given a word.

2.1. *Visual input*

The features of the images are extracted by training a 16-layer convolutional neural network (CNN) (Simonyan and Zisserman, 2014) on an object recognition task.[2] The network is trained to discriminate among 1000 different object labels on the ImageNet dataset (Deng *et al.*, 2009). The last layer of the CNN before the classification layer contains high level visual features of the images, invariant to particulars such as position, orientation or size. We use the activation vector from this layer as a representation of the visual scene described in the corresponding caption. Each caption is paired with such a 4096-dimensional vector and used as input to a cross-situational word learner. Figure 1 shows three sample images from the F8k dataset most closely aligned with a particular dimension, as measured by the cosine similarity between the image and a unit vector parallel to the dimension axis. For example, dimension 1000 seems to be related to water, 2000 to dogs or perhaps grass, and 3000 to children.

2. We used the F8k and F30k features available at `http://cs.stanford.edu/people/karpathy/deepimagesent/` and the data handling utilities from `https://github.com/karpathy/neuraltalk` for our experiments. The pre-trained CNN can be used through the Caffe framework (Jia *et al.*, 2014) and is available at the ModelZoo `https://github.com/BVLC/caffe/wiki/Model-Zoo`

Dimension	Top 3 images
1000	
2000	
3000	

Figure 1. Dimensions with three most closely aligned images from F8k.

2.2. *Learning algorithm*

We adapt the IBM model 1 estimation algorithm in the following ways[3]: (i) like Fazly *et al.* (2010) we run it in an online fashion, and (ii) instead of two sequences of words, our input consists of one sequence of words on one side, and a vector of real values representing the image on the other side. The dimensions are indexes into the visual feature "vocabulary", while the values are interpreted as weights of these "vocabulary items". In order to get an intuitive understanding of how the model treats the values in the feature vector, we could informally liken these weights to word counts. As an example consider the following input with a sentence and a vector of 5 dimensions (i.e. 5 features):

– The blue sky

– $(2, 0, 2, 1, 0)$

Our model treats this equivalently to the following input, with the values of the dimensions converted to "feature occurrences" of each feature f_n.

– The blue sky

– $f_1\ f_1\ f_3\ f_3\ f_4$

3. The source code for our model is available at `https://github.com/kadarakos/IBMVisual`

Short title 79

The actual values in the image vectors are always non-negative, since they come from a rectified linear (ReLu) activation. However, they can be fractional, and thus strictly speaking cannot be literal counts. We simply treat them as generalized, fractional feature "counts". The end result is that given the lists of words in the image descriptions and the corresponding image vectors the model learns a probability distribution $t(f|w)$ over feature-vector indexes f for every word w in the descriptions.

Algorithm 1 Sentence-vector alignment model (VISUAL)

1: **Input:** visual feature vectors paired with sentences $((V_1, S_1), \ldots, (V_N, S_N))$
2: **Output:** translation table $t(f|w)$
3: $D \leftarrow$ dimensionality of feature vectors
4: $\epsilon \leftarrow 1$ ▷ Smoothing coefficient
5: $a[f, w] \leftarrow 0, \forall f, w$ ▷ Initialize count tables
6: $a[\cdot, w] \leftarrow 0, \forall w$
7: $t(f|w) \leftarrow \frac{1}{D}$ ▷ Translation probability $t(f|w)$
8: **for** each input pair (vector V, sentence S) **do**
9:     **for** each feature index $f \in \{1, \ldots, D\}$ **do**
10:         $Z_f \leftarrow \sum_{w \in S} t(f|w)$ ▷ Normalization constant Z_f
11:         **for** each word w in sentence S **do**
12:             $c \leftarrow \frac{1}{Z_f} \times V[f] \times t(f|w)$ ▷ Expected count c
13:             $a[f, w] \leftarrow a[f, w] + c$
14:             $a[\cdot, w] \leftarrow a[\cdot, w] + c$ ▷ Updates to count tables
15:             $t(f|w) \leftarrow \frac{a[f,w]+\epsilon}{a[\cdot,w]+\epsilon D}$ ▷ Recompute translation probabilities

This is our sentence-vector alignment model, VISUAL. In the interest of cognitive plausibility, we train it using a single-pass, online algorithm. Algorithm 1 shows the pseudo-code. Our input is a sequence of pairs of D-dimensional feature vectors and sentences, and the output is a translation table $t(f|w)$. We maintain two count tables of expected counts $a[f, w]$ and $a[\cdot, w]$ which are used to incrementally recompute the translation probabilities $t(f|w)$. The initial translation probabilities are uniform (line 7). In lines 12-14 the count tables are updated, based on translation probabilities weighted by the feature value $V[f]$, and normalized over all the words in the sentence. In line 15 the translation table is in turn updated.

"7" not separate from "line"...

2.3. *Baseline models*

To asses the quality of the meaning representations learned by our sentence-vector alignment model VISUAL, we compare its performance in a set of tasks to the following baselines:

– MONOLING: instead of aligning each sentence with its corresponding visual vector, this variation aligns two copies of each sentence with each other, and thus

learns word representations based on word-word co-occurrences[4].

– WORD2VEC: for comparison we also report results with the skip-gram embedding model, also known as WORD2VEC which builds word representations based on word-word co-occurrences as well (Mikolov *et al.*, 2013a; Mikolov *et al.*, 2013b). WORD2VEC learns a vector representation (embedding) of a word which maximizes performance on predicting words in a small window around it.

3. Experiments

3.1. *Image datasets*

We use image-caption datasets for our experiments. F8k (Rashtchian *et al.*, 2010) consists of 8000 images and five captions for each image. F30k (Young *et al.*, 2014) extends the F8k and contains 31,783 images with five captions each summing up to 158,915 sentences. For both data sets we use the splits from Karpathy and Fei-Fei (2014), leaving out 1000 images for validation and 1000 for testing from each set. Table 1 summarizes the statistics of the Flickr image-caption datasets.

bold

italics

to be placed under data, please

Table 1. Flickr image caption datasets

	F8k	F30k
Train images	6,000	29,780
Validation images	1,000	1,000
Test images	1,000	1,000
Image in total	8,000	31,780
Captions per image	5	5
Captions in total	40,000	158,900

For the Single-concept image descriptions experiments reported in Section 3.4, we also use the ILSVRC2012 subset of ImageNet (Russakovsky *et al.*, 2014), a widely-used data set in the computer vision community. It is an image database that annotates the WordNet noun synset hierarchy with images. It contains 500 images per synset on average.

bdc

3.2. *Word similarity experiments*

A common evaluation task for assessing the quality of learned semantic vectors for words is measuring word similarity. A number of experiments have elicited human ratings on the similarity and/or relatedness of a list of word pairs. For instance one of the data sets we used was the SimLex999 data set, which contains similarity judgments

4. This model does not estimate probabilities of translation of a word to itself, that is probabilities of the form $t(w|w)$.

for 666 noun pairs (organ-liver 6.15), 222 verb pairs (occur-happen 1.38) and 111 adjective pairs (nice-cruel 0.67) elicited by 500 participants recruited from Mechanical Turk. These types of data sets are commonly used as benchmarks for models of distributional semantics, where the learned representations are expected to show a significant positive correlation with human similarity judgments on a large number of word pairs.

We selected a subset of the existing benchmarks according to the size of their word pairs that overlap with our restricted vocabulary. We ran a statistical power analysis test to estimate the minimum number of required word pairs needed in our experiments. The projected sample size was $N = 210$ with $p = .05$, effect-size $r = .2$ and $power = 0.9$. Thus some of the well-known benchmarks were excluded due to their small sample size after we excluded words not present in our datasets.[5]

The four standard benchmarks that contain the minimum number of word pairs are: the full WS-353 (Finkelstein *et al.*, 2001), MTurk-771 (Radinsky *et al.*, 2011), MEN (Bruni *et al.*, 2014) and SimLex999 (Hill *et al.*, 2014). Note that the MTurk dataset only contains similarity judgments for nouns. Also, a portion of the full WordSim-353 dataset reports relatedness ratings instead of word similarity.

3.3. *Effect of concreteness on similarity judgments*

The word similarity judgments provide a macro evaluation about the overall quality of the learned word representations. For more fine-grained analysis we turn to the dichotomy of concrete (e.g. *chair, car*) versus abstract (e.g. *love, sorrow*) nouns. Evidence presented by (Recchia and Jones, 2012) shows that in naming and lexical decision tasks the early activation of abstract concepts is facilitated by rich linguistic contexts, while physical contexts promote the activation of concrete concepts. Based on these recent findings, Bruni *et al.* (2014) suggest that in case of computational models *concrete* words (such as names for physical objects and visual properties) are easier to learn from perceptual/visual input and *abstract* words are mainly learned based on their co-occurrence with other words in text. Following Bruni *et al.* (2014), but using novel methodology, we also test this idea and examine whether more concrete words benefit more from visual features compared with less concrete ones.

In their work Bruni *et al.* (2014) use the automatic method from Turney *et al.* (2011) to assign concreteness values to words and split the MEN corpus in concrete and abstract chunks. From their experiments they draw the conclusion that visual information boosts their models' performance on concrete nouns. However, whereas the multi-modal embeddings of Bruni *et al.* (2014) are trained using an unbalanced corpus of large quantities of textual information and far poorer visual stimuli, our visual embeddings are learned on a parallel corpus of sentences paired with images.

5. These include RG-65 (Rubenstein and Goodenough, 1965), MC-30 (Miller and Charles, 1991) and YP-130 (Yang and Powers, 2006).

To our purposes, this balance in the sources of information is critical as we aim at modeling word learning in humans. As a consequence of this setting we rather hypothesized that solely relying on visual features would result in better performance on more concrete words than on abstract ones and conversely, learning language solely from textual features would lead to higher correlations on the more abstract portion of the vocabulary.

To test this hypothesis, MEN, MTurk and Simlex999 datasets were split in two halves based on concreteness score of the word pairs. The "abstract" and "concrete" subclasses for each data set are obtained by ordering the pairs according to their concreteness and then partition the ordered tuples in halves. We defined the concreteness of a word pair as the product of the concreteness scores of the two words. The scores are taken from the University of South Florida Free Association Norms dataset (Nelson *et al.*, 1998). Table 2 provides an overview of the benchmarks we use in this study. Column "Concreteness" shows the average concreteness scores of all words pairs per data set, while columns "Concrete" and "Abstract" contain the average concreteness of the concrete and abstract halves of the word-pairs respectively.

Bold 1. italics

Table 2: Summary of the word-similarity benchmarks, showing the number of word pairs in the benchmarks and the size of their overlap with the F8k and F30k data sets. The table also reports the average concreteness of the whole, concrete and abstract portions of the benchmarks.

	#Pairs			Concreteness		
	Total	F8k	F30k	Full set	Concrete	Abstract
WS353	353	104	232	25.09	35.44	16.22
SimLex999	999	412	733	23.86	35.72	11.99
MEN	3000	2069	2839	29.77	36.28	23.26
MTurk771	771	295	594	25.89	34.02	16.16

under please

3.4. *Word production*

Learning multi-modal word representations gives us the advantage of replicating real-life tasks such as naming visual entities. In this study, we simulate a word production task as follows: given an image from the test set, we rank all words in our vocabulary according to their cosine similarity to the visual vector representing the image. We evaluate these ranked lists in two different ways.

3.4.1. *Multi-word image descriptions.*

We use images from the test portion of the F8k and F30k datasets as benchmarks. These images are each labeled with up to five captions, or multi-word descriptions of the content of the image. To evaluate the performance of our model in producing words for each image, we construct the target description of an image as the union of

Short title 83

the words in all its captions (with stop-words[6] removed). We compare this set with the top N words in our predicted ranked word list. As a baseline for this experiment we implemented a simple frequency baseline FREQ, which for every image retrieves the top N most frequent words. The second model COSINE uses our VISUAL word-embeddings and ranks the words based on their cosine similarity to the given image. The final model PRIOR implements a probabilistic interpretation of the task

$$P(w_i|i_j) \propto P(i_j|w_i) \times P(w_i), \qquad [1]$$

where w_i is a word from the vocabulary of the captions and i_j is an image from the collections of images I. The probability of an image given a word is defined as

$$P(i_j|w_i) = \frac{\text{cosine}(i_j, w_i)}{\sum_{k=1}^{|I|} \text{cosine}(i_k, w_i),} \qquad [2]$$

where $\text{cosine}(i_j, w_i)$ is the cosine between the vectorial representation of i_j and the VISUAL word-embedding w_i. Since in any natural language corpus the distribution of word frequencies is expected to be very heavy tailed, in the model PRIOR, rather than using maximum likelihood estimation, we reduce the importance of the differences in word-frequencies and smooth the prior probability $P(w_i)$ as described by Equation 3 - where N is the number of words in the vocabulary.

bdc

$$P(w_i) = \frac{\log(count(w_i))}{\sum_{j=1}^{N} \log(count(w_j))} \qquad [3]$$

As a measure of performance, we report Precision at 5 (P@5) between the ranked word list and the target descriptions; i.e., proportion of correct target words among the top 5 predicted ranked words. Figure 2 shows an example of an image and its multi-word captions in the validation portion of the F30k dataset.

3.4.2. *Single-concept image descriptions.*

Even though we use separate portions of F8k and F30k for training and testing, these subsets are still very similar. To test how general the VISUAL word representations are we use images from the ILSVRC2012 subset of ImageNet (Russakovsky *et al.*, 2014) as benchmark. The major difference between these images and the ones from F8k and F30k datasets is that labels of the images in ImageNet are synsets from WordNet, which identify a single concept present in the image instead of providing a natural descriptions of its full content. Providing a quantitative evaluation in this case is not straightforward for two main reasons. First, the vocabulary of our model is restricted and the synsets in the ImageNet dataset are quite varied. Second, the

6. Function words such as *the, is, at, what, there* we used the stop-word list from the Python library NLTK.

L24 TAL. Volume 55 - nº 3/2014

~~Short title 13~~

bold

Figure 2. Multiword image description example. Below the image are the 5 captions describing the image, the union of words that we take as targets, the to 5 predicted and the list of correct words and the P@5 score for the given test case.

ok?

italics

A boy in a blue shirt and white helmet is riding a white bike.
A boy in blue is riding his bike in a skate park.
A boy on a BMX bike
A cyclist riding on their front wheel on the asphalt.
The man is on a black and white bike.

Descriptors: blue boy skate shirt asphalt helmet
park cyclist black bike wheel front
riding white bmx man

Predicted: bike bicycle riding man biker
Overlap: riding bike man
P@5: 0.6

if possible...

synset labels can be very precise, much more so than the descriptions provided in the captions that we use as our training data.

To attempt to solve the vocabulary mis-match problem, we use synset hypernyms from WordNet as substitute target descriptors. If none of the lemmas in the target synset are in the vocabulary of the model, the lemmas in the hypernym synset are taken as new targets, until we reach the root of the taxonomy. However, we find that in a large number of cases these hypernyms are unrealistically general given the image. Figure 3 illustrates these issues.

4. Results

We evaluate our model on two main tasks: simulating human judgments of word similarity [7] and producing labels for images. For all performance measures in this sections (Spearman's ρ, P@5) we estimated the confidence intervals using the Bias-corrected Accelerated bootstrapping method [8] (Efron, 1982).

7. We made available the source code used for running word similarity/relatedness experiments on `https://bitbucket.org/kadar_akos/wordsims`.

8. Provided by the scikits-bootstrap Python package `https://github.com/cgevans/scikits-bootstrap`.

Figure 3. Example of the Single-concept image description task from the validation portion of the ILSVRC2012 subset of ImageNet. The terms "sea anemone" and "anemone" are unknown to VISUAL and "animal" is the first word among it's hypernyms that appear in the vocabulary of F30k.

Label: sea anemone anemone
Hypernym: animal

4.1. *Word Similarity*

We simulate the word similarity judgment task using the induced word vectors by three models: VISUAL, MONOLING, and WORD2VEC. All models were trained on the tokenized training portion of the F30k data set. While VISUAL is presented with pairs of captions and the 4096 dimensional image-vectors, MONOLING and WORD2VEC[9] are trained solely on the sentences in the captions. The smoothing coefficient $\epsilon = 1.0$ was used for VISUAL and MONOLING. The WORD2VEC model was run for one iteration with default parameters except for the minimum word count (as our models also consider each word in each sentence): feature-vector-size=100, alpha=0.025, window-size=5, min-count=5, downsampling=False, alpha=0.0001, model=skip-gram, hierarchical-sampling=True, negative-sampling=False.

Figure 4 illustrates the correlation of the similarity judgments by the three models with those of humans on four datasets. Table 3 shows the results in full detail: it reports the Spearman rank-order correlation coefficient between the human similarity judgments and the pairwise cosine similarities of the word vectors per data set along with the confidence intervals estimated by using bootstrap (the correlation values marked by a * were significant at level $p < 0.05$).

Although VISUAL achieves a higher correlation than the other two models on all datasets, the overlapping confidence intervals suggest that in this particular setting, both VISUAL and WORD2VEC perform very similarly in approximating human sim-

9. we used the word2vec implementation from the gensim Pyhton package available at `https://radimrehurek.com/gensim/models/word2vec.html`.

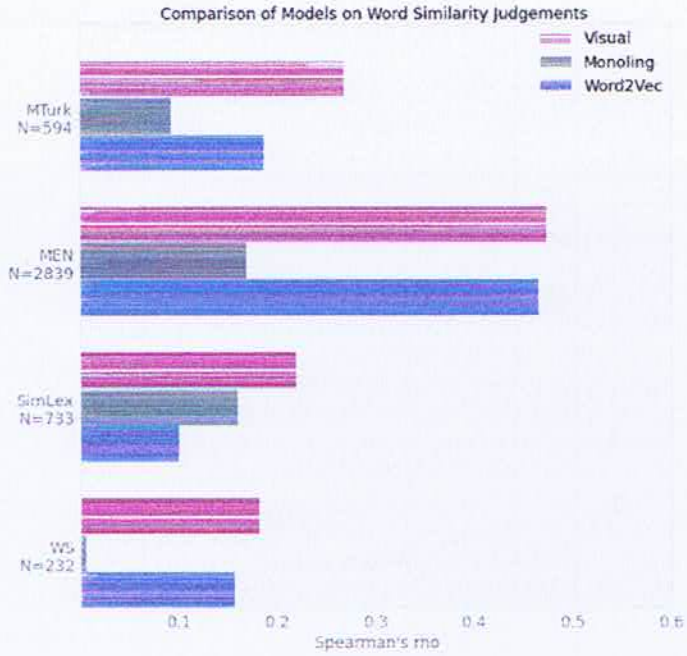


Figure 4: Comparison of models on approximating word similarity judgments. The length of the bars indicate the size of the correlation measured by Spearman's ρ, longer bars indicate better similarity between the models' predictions and the human data. The labels on the y-axis contain the names of the data sets and indicate the number of overlapping word pairs with the vocabulary of the F30k data set. All models were trained on the training portion of the F30k data set.

ilarity judgments. This result is particularly interesting as these models exploit different sources of information: The input to WORD2VEC is text only (i.e., the set of captions) and it learns from word-word co-occurrences, while VISUAL takes pairs of image vectors and sentences as input, and thus learns from word-scene co-occurrences.

The significant medium-sized correlation ($p < .001$, $\rho = 0.47$ 95% CI [0.44, 0.50]) with reasonably narrow confidence intervals on the large number of samples, $N = 2839$, of the MEN data set supports the hypothesis that the similarities between the meaning representations learned by VISUAL mirror the distance between word pairs as estimated by humans. This result suggests that it is feasible to learn word meanings from co-occurrences of sentences with noisy visual scenes. However, on all other data sets, the effect sizes for all models are small and their performances vary considerably given different subsamples of the benchmarks.

Table 3: Word similarity correlations with human judgments measured by Spearman's ρ. Models were trained on the training portion of the F30k data set. The * next to the values marks the significance of the correlation at level $p < 0.05$. The confidence intervals for the correlation are estimated using bootstrap.

	WS	SimLex	MEN	MTurk
VISUAL	**0.18***	**0.22***	**0.47***	**0.27***
	CI[0.05, 0.32]	CI[0.15, 0.29]	CI[0.44, 0.50]	CI[0.19, 0.34]
MONOLING	0.08	0.18*	0.23*	0.17*
	CI[-0.06, 0.21]	CI[0.11, 0.25]	CI[0.19, 0.26]	CI[0.04, 0.19]
WORD2VEC	0.16*	0.10*	0.47*	0.19*
	CI[0.02, 0.28]	CI[0.02, 0.17]	CI[0.43 0.49]	CI[0.11, 0.26]

4.1.1. *Concreteness*

Based on the previous findings of Bruni *et al.* (2014) we expected that models relying on perceptual cues perform better on the concrete portion of the word-pairs in the word-similarity benchmarks. Furthermore, we expected approximating human word similarity judgments on concrete word-pairs to be generally easier. As discussed in Section 3.3, we split the data sets into *abstract* and *concrete* halves and ran the word similarity experiments on the resulting portions of the word-pairs for comparison. Table 4 only reports the results on MEN and Simlex999 as these were the only benchmarks that had at least 200 word-pairs after partitioning. Table 2 summarizes the average concreteness of the different portions of the data sets.

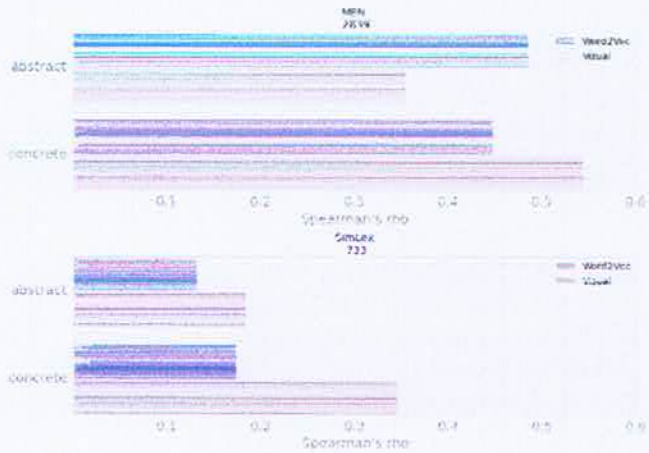
On all data sets VISUAL seems to perform considerably better on the concrete word-pairs then on abstract ones. On the abstract half of the MEN data set the performance of VISUAL is $\rho = 0.35$, 95% $CI[0.29, 0.41]$, while it is $\rho = 0.56$, 95% $CI[0.49, 0.59]$ on the concrete portion. The non-overlapping confidence intervals support the hypothesis that VISUAL does significantly better on the concrete word pairs. This pattern, however, is not observed for WORD2VEC as there is no significant difference in its performance given the different concreteness levels of the word pairs. Splitting the word pairs in two groups based on their concreteness scores reveals that performance of VISUAL is affected by concreteness and that it only performs better than WORD2VEC on the more concrete word pairs. Another pattern that the analysis reveals is that the average concreteness of the data sets is reflected in the performance of the models: for both VISUAL and WORD2VEC the rank of their performance follows the rank of concreteness of the benchmarks.

1. Bold

Table 4: The table reports the Spearman rank-order correlation coefficient on the abstract and concrete portions of the data sets separately as well as the confidence intervals around the effect-sizes estimated by using bootstrap. The * next to the values indicates significance at level $p < 0.05$.

italics

	MEN		SimLex	
	Abstract	Concrete	Abstract	Concrete
Visual	0.35*	0.55*	0.16*	0.39*
	CI[0.29, 0.41]	CI[0.49, 0.59]	CI[0.04, 0.25]	CI[0.28, 0.47]
Word2Vec	0.48	0.45	0.14	0.18
	CI[0.43, 0.53]	CI[0.39, 0.50]	CI[0.02, 0.25]	CI[0.07, 0.29]



Bold 1.

Figure 5: Models' performance on word similarity judgments as a function of the concreteness of the word pairs.

italics

4.2. *Word production*

In this set of experiments, we evaluate the word meaning vectors learned by VISUAL by simulating the task of word production for an image, as described in Section 3.4. These experiments can be viewed as computational simulations of a language task where human subjects associate words to given images. Words were ranked according to their cosine similarity to a given image vector. The VISUAL model was trained on the training portion of the F8k and F30k data sets. We report results on two variations of the word production task: multi-word image descriptors, and single-concept image descriptors.

bdc

4.2.1. *Multi-word image descriptors*

The objective of the model in this experiment is to rank only words in the top N that occur in the set containing all words from the concatenation of the 5 captions of a given image with stop-words removed. The ranking models used for these experiments (FREQ, COSINE, and PRIOR) are described in Section 3.4. Table 5 reports the results of the experiments on the respective test portions of the F8k and F30k datasets as estimated by P@5. We estimated the variability of the models' performance by calculating these measures per sample and estimating the confidence intervals around the means using bootstrap.

On these particular data sets the naive frequency baseline can perform particularly well: by only retrieving the sequence $< wearing, woman, people, shirt, blue >$, the ranking model FREQ scores a P@5=.27 on F30k. Incorporating both the meaning representations learned by VISUAL and the prior probabilities of the words, the non-overlapping confidence intervals suggest that PRIOR significantly outperforms FREQ P@5=0.42, 95% $CI[0.41, 0.44]$.

In addition to P@5, we also report the number of word types that were retrieved correctly given the images (column Words@5 on Table 5). This measure was inspired by the observation that by focusing only on the precision scores it seems like incorporating visual information rather than just using raw word-frequency statistics provides a significant, but small advantage. However, taking into consideration that PRIOR retrieves 178 word types correctly, suggests that it can retrieve less generic words that are especially descriptive of fewer scenes.

To have a more intuitive grasp on the performance of PRIOR it is worth taking also into consideration the distribution of P@5 scores over the test cases. When trained and tested on F30k in most cases (34%) PRIOR retrieves two words correctly in the top 5 and in 23% and 25% of the cases it retrieves one and three respectively. In only 6% of the time $P@5 = 0$, which means that it is very unlikely that PRIOR named unrelated concepts given an image. These results suggest that VISUAL learns word meanings that allow for labeling unseen images with reasonable accuracy using a large variety of words.

4.2.2. *Single-concept image descriptors*

The motivation for this experiment was to assess the generalizability of the word-representations learned by VISUAL. Similarly to the previous task, the goal here is to associate words to a given image, but in this case the images are drawn from the validation set of ILSVRC2012 portion of ImageNet (Russakovsky *et al.*, 2014). Providing quantitative results is not as straightforward as in the case of multi-word image descriptors, since these images are not labeled with target descriptions, but with a synset from WordNet. As demonstrated in Figure 6, some of the lemmas in the target synsets are far too specific or unnatural for our purposes, for example *schooner* for an image depicting a sailboat or *alp* for an image of a mountain. In other cases, a partic-

bold 1. bdc italics

Table 5: Results for the Multi-word image descriptors experiments reported on the test sets of F8k and F30k. Words@5 the number of correctly retrieved word types in the top 5. The confidence intervals below P@5 scores were estimated using bootstrap.

	F8k		F30k	
	P@5	Words@5	P@5	Words@5
FREQ	0.20	5	0.27	5
	CI[0.19, 0.21]		CI[0.26, 0.29]	
COSINE	0.16	310	0.14	371
	CI[0.15, 0.17]		CI[0.13, 0.15]	
PRIOR	**0.44**	135	**0.42**	178
	CI[0.42, 0.45]		CI[0.41, 0.44]	

ular object is named which might not be the most salient one, for example *freight car* for a picture of a graffiti with three pine trees on the side of railway carriage.

bdc

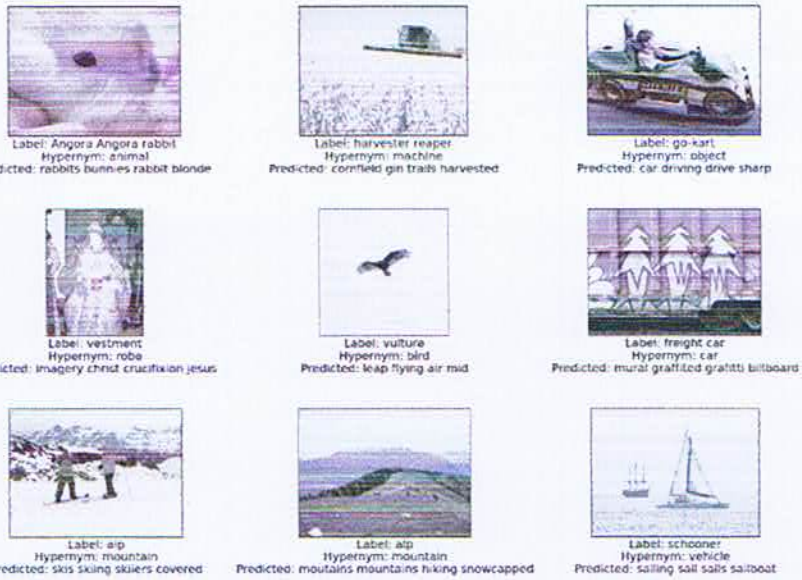
We made an attempt to search through the lemmas in the hypernym paths of the synsets until a known target lemma is reached. However, as demonstrated by examples in Figure 6, these hypernyms are often very general (e.g. *device*) and predicting such high-level concepts as descriptors of the image is unrealistic. In other cases, the lemmas from the hypernym synsets are simply misleading; for example, *wood* for describing a wooden wind instrument. As can be seen in the examples in Figure 6, the top ranked words predicted by our model are in fact conceptually more similar to the images covering a variety of objects and concepts than the labels specified in the dataset. bdc

We conclude that in the future to quantitatively investigate the cognitive plausibility of cross-situational models of word learning, the collection of feature production norms for ImageNet (Russakovsky *et al.*, 2014) would be largely beneficial. , λ

5. Discussion and conclusion

We have presented a computational cross-situational word learning model that learns word meanings from pairs images and their natural language descriptions. Unlike previous word learning studies which often rely on artificially generated perceptual input, the visual features we extract from images of natural scenes offers a more realistic simulation of the cognitive tasks humans face, since our data includes a significant level of ambiguity and referential uncertainty.

Our results suggest that the proposed model can learn meaningful representations for individual words from varied scenes and their multiword descriptions. Learning



Bold /. Figure 6: The caption above the images show the target labels, the hypernyms that were considered as a new target if the original was not in the vocabulary and the top N predicted words. In a large number of cases the guesses of the model are conceptually similar to the images, although, do not actually overlap with the labels or the hypernyms. italics

takes place incrementally and without assuming access to single-word unambiguous utterances or corrective feedback. When using the learned visual vector representations for simulating human ratings of word-pair similarity, our model shows significant correlation with human similarity judgments on a number of benchmarks. Moreover, it moderately outperforms other models that only rely on word-word co-occurrence statistics to learn word meaning.

The comparable performance of visual versus word-based models seems to be in line with Louwerse (2011), who argues that linguistic and perceptual information show a strong correlation and therefore meaning representations solely based on linguistic data are not distinguishable from representations learned from perceptual information. However, an analysis of the impact of word concreteness on the performance of our model shows that visual features are especially useful when estimating the similarity of more concrete word pairs. In contrast, models that rely on word-based cues do not show such improvement when judging the similarity of concrete word pairs. These results suggest that these two sources of information might best be viewed as complementary, as also argued by Bruni *et al.* (2014).

We also used the word meaning representations that our model learns from visual input to predict the best label for a given image. This task is similar to word production in language learners. Our quantitative and qualitative analyses show that the learned representations are informative and the model can produce intuitive labels for the images in our dataset. However, as discussed in the previous section, the available image collections and their labels are not developed to suit our purpose, as most of the ImageNet labels are too detailed and at a taxonomic level which is not compatible with how language learners name a visual concept.

Finally, a natural next step for this model is to also take into account cues from sentence structure. For example, Alishahi and Chrupała (2012) try to include basic syntactic structure by introducing a separate category learning module into their model. Alternatively, learning sequential structure and visual features could be modeled in an integrated rather than modular fashion, as done by the multimodal captioning systems based on recurrent neural nets (see Section 1.2). We are currently developing this style of integrated model to investigate the impact of structure on word learning from a cognitive point of view.

6. References

Alishahi A., Chrupała G., "Concurrent acquisition of word meaning and lexical categories", *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, p. 643-654, 2012.

Beekhuizen B., Fazly A., Nematzadeh A., Stevenson S., "Word learning in the wild: What natural data can tell us", *Proceedings of the 35th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society*, 2013.

Brown P. F., Pietra V. J. D., Pietra S. A. D., Mercer R. L., "The mathematics of statistical machine translation: Parameter estimation", *Computational linguistics*, vol. 19, nº 2, p. 263-311, 1993.

Bruni E., Tran N.-K., Baroni M., "Multimodal Distributional Semantics", *J. Artif. Intell. Res. (JAIR)*, vol. 49, p. 1-47, 2014.

Carey S., "The Child as Word Learner", *in* M. Halle, J. Bresnan, G. A. Miller (eds), *Linguistic Theory and Psychological Reality*, The MIT Press, 1978.

Chen X., Zitnick C. L., "Learning a Recurrent Visual Representation for Image Caption Generation", *arXiv preprint arXiv:1411.5654*, 2014.

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., "Imagenet: A large-scale hierarchical image database", *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, p. 248-255, 2009.

Donahue J., Hendricks L. A., Guadarrama S., Rohrbach M., Venugopalan S., Saenko K., Darrell T., "Long-term recurrent convolutional networks for visual recognition and description", *arXiv preprint arXiv:1411.4389*, 2014.

Efron B., *The jackknife, the bootstrap and other resampling plans*, vol. 38, SIAM, 1982.

Fang H., Gupta S., Iandola F., Srivastava R., Deng L., Dollár P., Gao J., He X., Mitchell M., Platt J. *et al.*, "From captions to visual concepts and back", *arXiv preprint arXiv:1411.4952*, 2014.

Fazly A., Alishahi A., Stevenson S., "A probabilistic computational model of cross-situational word learning", *Cognitive Science: A Multidisciplinary Journal*, vol. 34, n° 6, p. 1017-1063, 2010.

Fellbaum C., *WordNet*, Wiley Online Library, 1998.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G., Ruppin E., "Placing search in context: The concept revisited", *Proceedings of the 10th international conference on World Wide Web*, ACM, p. 406-414, 2001.

Fleischman M., Roy D., "Intentional Context in Situated Language Learning", *Ninth Conference on Computational Natural Language Learning*, 2005.

Frank M. C., Goodman N. D., Tenenbaum J. B., "A Bayesian Framework for Cross-Situational Word-Learning", *Advances in Neural Information Processing Systems*, vol. 20, 2007.

Hill F., Reichart R., Korhonen A., "Simlex-999: Evaluating semantic models with (genuine) similarity estimation", *arXiv preprint arXiv:1408.3456*, 2014.

Hochreiter S., Schmidhuber J., "Long short-term memory", *Neural computation*, vol. 9, n° 8, p. 1735-1780, 1997.

Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S., Darrell T., "Caffe: Convolutional Architecture for Fast Feature Embedding", *arXiv preprint arXiv:1408.5093*, 2014.

Karpathy A., Fei-Fei L., "Deep visual-semantic alignments for generating image descriptions", *arXiv preprint arXiv:1412.2306*, 2014.

Kiros R., Salakhutdinov R., Zemel R. S., "Unifying visual-semantic embeddings with multimodal neural language models", *arXiv preprint arXiv:1411.2539*, 2014.

Lazaridou A., Pham N. T., Baroni M., "Combining Language and Vision with a Multimodal Skip-gram Model", *arXiv preprint arXiv:1501.02598*, 2015.

Li P., Farkas I., MacWhinney B., "Early Lexical Development in a Self-organizing Neural Network", *Neural Networks*, vol. 17, p. 1345-1362, 2004.

Louwerse M. M., "Symbol interdependency in symbolic and embodied cognition", *Topics in Cognitive Science*, vol. 3, n° 2, p. 273-302, 2011.

MacWhinney B., *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*, Psychology Press, 2014.

Mao J., Xu W., Yang Y., Wang J., Yuille A. L., "Explain images with multimodal recurrent neural networks", *arXiv preprint arXiv:1410.1090*, 2014.

Matusevych Y., Alishahi A., Vogt P., "Automatic generation of naturalistic child–adult interaction data", *Proceedings of the 35th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society*, p. 2996-3001, 2013.

Mikolov T., Chen K., Corrado G., Dean J., "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems*, p. 3111-3119, 2013b.

Miller G. A., Charles W. G., "Contextual correlates of semantic similarity", *Language and cognitive processes*, vol. 6, n° 1, p. 1-28, 1991.

Nelson D., McEvoy C., Schreiber T., "The University of South Florida word association, rhyme, and word fragment norms. 1998 http://www. usf. edu", *FreeAssociation.[PubMed]*, 1998.

Pinker S., *Learnability and Cognition: The Acquisition of Argument Structure*, Cambridge, MA: MIT Press, 1989.

Quine W., *Word and Object*, Cambridge University Press, Cambridge, MA, 1960.

Radinsky K., Agichtein E., Gabrilovich E., Markovitch S., "A word at a time: computing word relatedness using temporal semantic analysis", *Proceedings of the 20th international conference on World wide web*, ACM, p. 337-346, 2011.

Rashtchian C., Young P., Hodosh M., Hockenmaier J., "Collecting image annotations using Amazon's Mechanical Turk", *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics, p. 139-147, 2010.

Recchia G., Jones M. N., "The semantic richness of abstract concepts", *Frontiers in human neuroscience*, 2012.

Regier T., "The Emergence of Words: Attentional Learning in Form and Meaning", *Cognitive Science: A Multidisciplinary Journal*, vol. 29, p. 819-865, 2005.

Rubenstein H., Goodenough J. B., "Contextual correlates of synonymy", *Communications of the ACM*, vol. 8, n° 10, p. 627-633, 1965.

Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A. C., Fei-Fei L., "ImageNet Large Scale Visual Recognition Challenge", *arXiv preprint arXiv:1409.0575*, 2014.

Simonyan K., Zisserman A., "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

Siskind J. M., "A computational study of cross-situational techniques for learning word-to-meaning mappings", *Cognition*, vol. 61, n° 1-2, p. 39-91, 1996.

Skocaj D., Kristan M., Vrecko A., Mahnic M., Janicek M., Kruijff G.-J. M., Hanheide M., Hawes N., Keller T., Zillich M. *et al.*, "A system for interactive learning in dialogue with a tutor", *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, IEEE, p. 3387-3394, 2011.

Smith L. B., Yu C., "Infants Rapidly Learn Word-Referent Mappings via Cross-Situational Statistics", *Cognition*, vol. 106, n° 3, p. 1558-1568, 2008.

Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A., "Going deeper with convolutions", *arXiv preprint arXiv:1409.4842*, 2014.

Turney P. D., Neuman Y., Assaf D., Cohen Y., "Literal and metaphorical sense identification through concrete and abstract context", *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, p. 680-690, 2011.

Venugopalan S., Xu H., Donahue J., Rohrbach M., Mooney R., Saenko K., "Translating Videos to Natural Language Using Deep Recurrent Neural Networks", *arXiv preprint arXiv:1412.4729*, 2014.

Vinyals O., Toshev A., Bengio S., Erhan D., "Show and tell: A neural image caption generator", *arXiv preprint arXiv:1411.4555*, 2014.

H 5

~~24 R° soumission à TAL 55-3~~

Short title 95

Vouloumanos A., "Fine-grained sensitivity to statistical information in adult word learning", *Cognition*, vol. 107, p. 729-742, 2008.

Vouloumanos A., Werker J. F., "Infants' learning of novel words in a stochastic environment", *Developmental Psychology*, vol. 45, p. 1611-1617, 2009.

Yang D., Powers D. M., *Verb similarity on the taxonomy of WordNet*, Citeseer, 2006.

Young P., Lai A., Hodosh M., Hockenmaier J., "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", *Transactions of the Association for Computational Linguistics*, vol. 2, p. 67-78, 2014.

Yu C., "A Statistical Associative Account of Vocabulary Growth in Early Word Learning", *Language Learning and Development*, vol. 4, n° 1, p. 32-62, 2008.

Yu C., Ballard D. H., "A unified model of early word learning: Integrating statistical and social cues", *Neurocomputing*, vol. 70, n° 13, p. 2149-2165, 2007.

Yu C., Smith L. B., "Rapid Word Learning under Uncertainty via Cross-Situational Statistics", *Psychological Science*, vol. 18, n° 5, p. 414, 2007.