End-to-end Image Captioning Exploits Multimodal Distributional Similarity

Pranava Madhyastha
p.madhyastha@sheffield.ac.uk
Josiah Wang
j.k.wang@sheffield.ac.uk
Lucia Specia
l.specia@sheffield.ac.uk

Department of Computer Science
The University of Sheffield
Sheffield, UK

Abstract

We hypothesize that end-to-end neural image captioning systems work seemingly well because they exploit and learn 'distributional similarity' in a multimodal feature space by mapping a test image to similar training images in this space and generating a caption from the same space. To validate our hypothesis, we focus on the 'image' side of image captioning, and vary the input image representation but keep the RNN text generation component of a CNN-RNN model constant. Our analysis indicates that image captioning models (i) are capable of separating structure from noisy input representations; (ii) suffer virtually no significant performance loss when a high dimensional representation is compressed to a lower dimensional space; (iii) cluster images with similar visual and linguistic information together. Our findings indicate that our distributional similarity hypothesis holds. We conclude that regardless of the image representation used image captioning systems seem to match images and generate captions in a learned joint image-text semantic subspace.

1 Introduction

Image description generation, or image captioning (IC), is the task of automatically generating a textual description for a given image. The generated text is expected to describe, in a single sentence, what is visually depicted in the image, for example the entities/objects present in the image, their attributes, the actions/activities performed, entity/object interactions (including quantification), the location/scene, etc. (e.g. "*a man riding a bike on the street*"). Significant progress has been made with *end-to-end* approaches to tackling this problem, where parallel image–description datasets such as Flickr30k [34] and MSCOCO [3] are used to train a CNN-RNN based neural network IC system [15, 28, 31]. Such systems have demonstrated impressive performance in the COCO captioning challenge[1] according to automatic metrics, seemingly even surpassing human performance in many instances (e.g. CIDEr score > 1.0 vs. human's 0.85) [3]. However, in reality, the performance of end-to-end systems is still far from satisfactory according to metrics based on human judgement[2]. Thus, despite the progress, this task is currently far from being a solved problem.

© 2018. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

[1] http://cocodataset.org/#captions-challenge2015

[2] http://cocodataset.org/#captions-leaderboard

In this paper, we challenge the common assumption that end-to-end IC systems are able to achieve strong performance because they have learned to 'understand' and infer semantic information from visual representations, i.e. they can for example deduce that "*a boy is playing football*" purely by learning directly from mid-level image features and the corresponding textual descriptions in an implicit manner, without explicitly modeling the presence of *boy*, *ball*, *green field*, etc. in the image. It is believed that the IC system has managed to infer that the phrase *green field* is associated with some 'green-like' area in the image and is thus generated in the output description, or that the word *boy* is generated because of some CNN activations corresponding to a young person. However, there seems to be no concrete evidence that this is the case. Instead, we hypothesize that the apparently strong performance of end-to-end systems is attributed to the fact that they exploit the *distributional similarity* property in a multimodal feature space. To our best knowledge, our paper gives the first empirical analysis on visual representations for the task of image captioning.

What we mean by 'distributional similarity' is that IC systems essentially attempt to find images from the training set that are most similar to a test image, and generate a caption from the most similar training instances (or generate a 'novel' description from a combination of training instances, for example by 'averaging' the descriptions). Previous work has alluded to this observation [14, 28], but it has not been thoroughly studied. This phenomenon could also be in part attributed to the fact that the datasets are repetitive and simplistic, with an almost constant and predictable linguistic structure [5, 16, 28]. Thus, while IC systems perform very well at the task of matching images to captions at surface-level, they do not truly understand images or language and use this understanding to generate image descriptions. Such misconception can deter true progress in the field. This paper aims to draw attention to this issue and to the importance of understanding how IC systems work and with that support work towards progress in the field that goes beyond optimizing for metrics to achieve state-of-the-art performance.

It is worth noting that we are interested in demonstrating the phenomenon of distributional similarity in IC, rather than achieving or improving state-of-the-art performance. As such, we do not resort to fine-tuning or extensive hyperparameter optimization or ensembles. Therefore, our model is not comparable to state-of-the-art models such as Vinyals et al. [28], which optimize IC by fine-tuning the image representations, exploring beam size, scheduled sampling, and using ensemble models. Instead, we vary only the image representation to demonstrate that end-to-end IC systems utilize distributional similarity on the image side to generate captions, regardless of the image representation used.

Our main contributions are:

(a) **An IC experiment** where we vary the input image representation but keep the RNN text generation model component constant (Section 3). This experiment demonstrates that regardless of the image representation (a continuous image embedding or a sparse, low-dimensional vector), end-to-end IC systems seem to utilize a visual-semantic subspace for IC.

(b) The introduction of **pseudo-random vectors** derived from object-level representations as a means to evaluate IC systems. Our results show that end-to-end models in this framework are remarkably capable of separating structure from noisy input representations.

(c) An experiment where IC models are conditioned on image representations **factorized and compressed to a lower dimensional space** (Section 4.1). We show that high dimensional image embeddings that are factorized to a lower dimensional representation and used as input to an IC model result in virtually no significant loss in performance, further strengthening our claim that IC models perform similarity matching rather than image understanding.

(d) An **analysis of different image representations and their transformed representations** (Section 4.2). We visualize the initial visual subspace and the learned joint visual semantic subspace and observe that the visual semantic subspace has learned to cluster images with similar visual and linguistic information together, further validating our claims of distributional similarity.

(e) An experiment where the IC model is **tested on an out-of-domain dataset** (Section 4.3), which has a slightly different image distribution. We observe that models show better performance on test sets that have a similar distribution as the training. Their performance deteriorates when the distributions are even slightly different.

Overall, our study demonstrates that end-to-end IC models implicitly learn and exploit multimodal similarity spaces rather than performing actual image understanding.

2 Model setting

For the experiments in Section 3, we base our implementation on the end-to-end approach by Karpathy and Fei-Fei [15]. We use the LSTM [12] based language model as described in Zaremba et al. [33], which is conditioned on the image information. For that, we first perform a linear projection of the image representation followed by a non-linearity:

$$Im_{feat} = \sigma(W \cdot I_m) \tag{1}$$

Here, $I_m \in \mathcal{R}^d$ is the d-dimensional initial image representation, $W \in \mathcal{R}^{n \times d}$ is the linear transformation matrix, σ is the non-linearity. We use Exponential Linear Units [4] as the non-linear activation in all our experiments. Following Vinyals et al. [27], we initialize the LSTM based caption generator with the projected image feature.

Training and Inference The image caption generator is trained to generate sentences conditioned on the image representation by minimizing a cross-entropy loss, i.e., the sentence-level loss corresponds to the sum of the negative log likelihood of the correct word being generated at each time step:

$$\Pr(S|Im_{feat};\theta) = \sum_t \log(\Pr(w_t|w_{t-1}..w_0;Im_{feat})) \tag{2}$$

where $\Pr(S|Im_{feat};\theta)$ is the sentence-level loss conditioned on the image feature Im_{feat} and $\Pr(w_t)$ is the probability of the word at time step t. This is trained with standard teacher forcing as described in Sutskever et al. [25] where the correct word information is fed to the next state in the LSTM.

Inference is typically performed with approximation techniques like beam search or sampling [15, 27]. In this paper, as we are mainly interested in studying the effect of different image representations, we focus on the language output that the models can most confidently produce. Therefore, unless stated otherwise we generate captions using a greedy argmax approach.

3 Image captioning with different image representations

In this section, we verify our hypothesis that a ‘distributional similarity’ space exists in end-to-end IC systems. Such systems attempt to match image representations in order to condition the RNN decoder to generate captions that are similar to the closest images, rather than actually understanding the image in order to describe it. We keep the IC model constant (Section 2) across experiments and vary only the image representation used. The different representations we experimented with are described in what follows.

3.1 Lower-bound image representation

Random: We condition the LSTM on a 300-dimensional vector comprising random values sampled uniformly between $[0,1)$[3]. This feature essentially gives us a worst-case image feature and thus provides an artificial lower bound.

3.2 Representations from image-level classification

We compare two CNNs – *VGG19* [24] and *ResNet152* [11] – both pre-trained on the ILSVRC challenge data [23]. We explore various representations derived from these CNNs:

Penultimate layer (*Penultimate*): Most previous attempts to IC use the output of the penultimate layer of a CNN pre-trained on ILSVRC. Previous work motivates using ‘off-the-shelf’ feature extractors in the framework of transfer learning [6, 20]. Such features have often been applied to image captioning [7, 9, 15, 18, 27, 31] and have been shown to produce state-of-the-art results. Therefore, we extract the ***fc7*** layer from *VGG19* ($4,096D$) and the ***pool5*** layer from *ResNet152* ($2,048D$) for each image.

Class prediction vector (*Softmax*): We also investigate higher-level image representations where each element in the vector is the estimated posterior probability of an object category appearing in that image. Note that the categories may not directly correspond to the captions in the dataset. While there are alternative methods that fine-tune the image network on a new set of object classes extracted in ways that are directly relevant to the captions [8, 30], we study the impact of off-the-shelf prediction vectors on the IC task. The intuition is that category predictions from pre-trained CNN classifiers may also be beneficial for IC, alongside the standard approach of using mid-level features from the penultimate layer. Therefore, for each image, we use the predicted category posterior distributions of *VGG19* and *ResNet152* for $1,000$ object categories.

Object class word embeddings (*Top-k*): Here we experiment with a method that utilizes the averaged word representations of top-k predicted object classes. We first obtain *Softmax* predictions using *ResNet152* for $1,000$ object categories (synsets) per image. We then select the objects that have a posterior probability score $> 5\%$ and use the 300-dimensional pre-trained word2vec [19] representations[4] to obtain the averaged vector over all retained object categories. This is motivated by the observation that averaged word embeddings can represent semantic-level properties and are useful for classification tasks [2].

[3] We also tried using 1,000-dimensions, which yielded similar but slightly poorer results.

[4] https://code.google.com/archive/p/word2vec/

3.3 Representations from object-level detections

We also explore representing images using information from object *detectors* that identify *instances* of object categories present in an image, rather than a global, image-level classification. This can potentially provide for a richer and more informative image representation. For this we use:

- *ground truth* (***Gold***) region annotations for instances of 80 pre-defined categories provided with MSCOCO. It is worth noting that these were annotated independently of the image captions, i.e. people writing the captions had no knowledge of the 80 categories. As such, there is no direct correspondence between the region annotations and image captions.

- a state-to-the-art object detector *YOLO* [21], pre-trained on MSCOCO for 80 categories (***YOLO-Coco***), and on MSCOCO and ILSVRC for over 9,000 categories (***YOLO-9k***). We use *YOLOv2*.

We explore several representations derived from instance-level object class annotations or detectors above:

Bag of objects (*BOO*): We represent each image as a sparse 'bag of objects' vector, where each element represents the frequency of occurrence for each object category in the image (***Counts***). We also explore an alternative representation where we only encode the presence or absence of the object category regardless of its frequency (***Binary***) to determine whether it is important to encode object counts in the image. These representations help us examine the importance of explicit object categories and in a sense interactions between object categories (e.g. *dog* and *ball*) in the image representation. We investigate whether such a sparse and high-level *BOO* representation is actually sufficient for generating image captions. It is also worth noting that *BOO* is different from the *Softmax* representation above as it encodes the *number* of object occurrences, not the *confidence* of class predictions at image level. We compare *BOO* representations derived from the ***Gold*** annotations (***Gold-Binary*** and ***Gold-Counts***) and both ***YOLO-Coco*** and ***YOLO-9k*** detectors (***Counts*** only).

Pseudo-random vectors: To further probe the capacity of the model to discern image representations in an image distributional similarity space, we propose a novel experiment in which we examine a type of representation where *similar images are represented using similar random vectors*, which we term as *pseudo-random vectors*. We form this representation from ***BOO Gold-Counts*** and ***BOO Gold-Binary***. More specifically, $Im_{feat} = \sum_{o \in \text{Objects}} f \times \phi_o$, where $\phi_o \in \mathcal{R}^d$ is an object-specific random vector and f is a scalar representing counts of the object category. In the case of ***Pseudorandom-Counts***, f is the frequency counts from ***Gold-Counts***. In the case of ***Pseudorandom-Binary***, f is either 0 or 1 based on ***Gold-Binary***. We use $d = 120$ for these experiments. Intuitively, these pseudo-random vectors appear random and noisy in the representational space as a result of the composition of (random) object category vectors, more specifically the multiplication of object category vectors by their frequency of occurrence and the addition of vectors across multiple object categories. We use these vectors to demonstrate that end-to-end IC models are capable of separating structure from noise, and thus exploit the distributional similarity property in a multimodal feature space.

3.4 Datasets and experimental setup

Dataset We evaluate image captioning conditioned on different representations on the most widely used dataset for IC, MSCOCO [3]. The dataset consists of 82,783 images for training, with at least five captions per image, totaling to 413,915 captions. We perform model selection on a 5000-image development set and report the results on a 5000-image test set using standard, publicly available splits[5] of the MSCOCO validation dataset as in previous work [15].

3.5 Image captioning results

We report results of IC on MSCOCO in Table 1, where the IC model (Section 2) is conditioned on the various image representations described in Section 3.1. As expected, using random image embeddings clearly does not provide any useful information and performs poorly. The CNN *softmax* representations with the same set of 1,000 object classes (*VGG19* and *ResNet152*) have very similar performance. We note that the posterior distribution may not directly correspond to words in the captions, i.e. many words and concepts are not contained in the set of object classes. Our results differ from those by Wu et al. [30] and Yao et al. [32] where the object classes have been fine-tuned to correspond directly to the caption vocabulary.

	Representation	B-1	B-2	B-3	B-4	M	C	S
	Random	0.48	0.24	0.11	0.07	0.11	0.07	0.03
Softmax	VGG19	0.62	0.43	0.29	0.19	0.20	0.61	0.13
	ResNet152	0.62	0.43	0.29	0.19	0.20	0.62	0.12
Penultimate	VGG19 (fc7)	0.65	0.46	0.32	0.22	0.21	0.69	0.14
	ResNet152 (pool5)	0.66	0.48	0.33	0.23	0.22	0.74	0.15
Embeddings	Top-*k*	0.62	0.42	0.28	0.19	0.20	0.63	0.13
BOO	Gold-Binary	0.65	0.47	0.32	0.22	0.22	0.75	0.15
	Gold-Counts	0.67	0.48	0.33	0.23	0.22	0.81	0.16
	YOLO-Coco	0.65	0.46	0.32	0.22	0.22	0.75	0.15
	YOLO-9k	0.64	0.45	0.31	0.21	0.20	0.68	0.13
Pseudo-random	Pseudorandom-Binary	0.65	0.46	0.31	0.21	0.21	0.73	0.14
	Pseudorandom-Counts	0.67	0.48	0.34	0.23	0.22	0.80	0.15

Table 1: Results on the MSCOCO test split, where we vary only the image representation and keep other parameters constant. The captions are generated with *beam* $= 1$. We report **B**LEU (1-4), **M**eteor, **C**IDEr and **S**PICE scores.

The performance of the *pool5* image representations shows a similar trend for *VGG19* and *ResNet152*, with *ResNet152* achieving slightly better scores than *VGG19*. We posit that the representations from the image network trained on object classes are able to capture more fine-grained image details.

The performance of the averaged top-*k* word embeddings is similar to that of the *Softmax* representation. This is interesting, since the averaged word representational information is mostly noisy: we combine top-*k* synset-level information into one single vector; however, it still performs competitively.

The performance of the *BOO* sparse 80-dimensional annotation vector is better than all other image representations judging by the CIDEr score. We note again that this occurs despite the fact that the annotations may not directly correspond to the semantic information in

[5] http://cs.stanford.edu/people/karpathy/deepimagesent

Method	B-1	B-2	B-3	B-4	M	C	S
PCA	0.66	0.48	0.34	0.24	0.22	0.75	0.15
ICA	0.66	0.48	0.34	0.24	0.22	0.74	0.15
PPCA	0.66	0.48	0.34	0.24	0.22	0.76	0.15
FULL	0.66	0.48	0.33	0.23	0.22	0.74	0.15

Table 2: Performance of compressed Pool5 representations.

Model	B-1	B-2	B-3	B-4	M	C
Pool5	0.60	0.41	0.26	0.17	0.14	0.29
SC	0.62	0.42	0.28	0.18	0.17	0.35
TDBU	0.60	0.40	0.26	0.17	0.17	0.34

Table 3: Performance of models on Flickr30k.

the captions or the images. The sparse representational information is indicative of the presence of only a subset of potentially useful objects. We notice two distinct patterns, a marked difference with *Binary* and *Count* representations. This takes us back to the motivation that image captioning requires information about objects, as well as interactions between objects and their attributes. Although our representation is really sparse on the object interactions, it captures the basic concept of the presence of more than one object of the same kind, and thus provides extra information. A similar trend was observed by Wang et al. [29], who further explored encoding the geometric and size information of objects into the representation, and by Yin and Ordonez [33], who learn interactions using a specified object-layout RNN.

We also notice that using predicted objects using *YOLOCoco* performs better than using *YOLO9k*. This is probably expected as *YOLOCoco* was trained on the same dataset hence producing better object proposals. We also observed that *YOLO9k* had a significant number of objects predicted for the test images that had not been seen in the training set (around 20%).

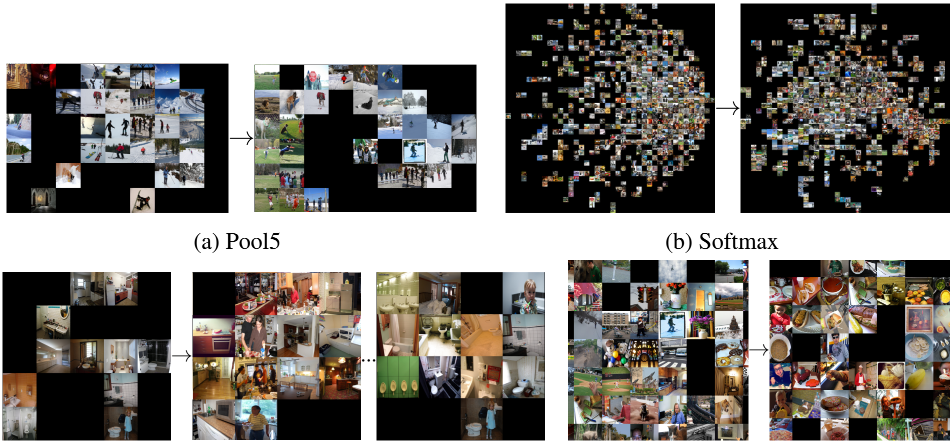
The most surprising result is the performance of the pseudo-random vectors. We notice that both the *pseudo-random binary* and *pseudo-random count* vectors perform almost as well as the *Gold* objects. This suggests that the conditioned RNN is able to remove noise and learn some sort of a common 'visual-linguistic' semantic subspace.

4 Analysis of distributional similarity in IC

In what follows we present further analyses on the different image representations to gain a better understanding of such representations and demonstrate our distributional similarity hypothesis.

4.1 Factorizing representations

In Section 3.5 we observed encouraging results from the bag of objects representation despite it being sparse, low-dimensional, and only partially relevant to captions. Interestingly, using pseudo-random vectors derived from a bag of objects also resulted in good performance despite the added noise. This leads to the question: are high-dimensional vectors necessary or relevant? To answer this question, we evaluate whether the performance of the model is significantly poorer if we reduce the dimensionality of the initial high dimensional representation. We experiment with three exploratory factor analysis-based methods – Principal Component Analysis (PCA) [10], Probabilistic Principal Component Analysis (PPCA) [26] and Independent Component Analysis (ICA) [13]. In all cases, we obtain 80-dimensional factorized representations from *ResNet152 pool5* (2048*D*), which is commonly used in IC. We summarize our results in Table 2. We observe that the representations obtained by all of the factored models seem to retain the necessary representational power to produce appropriate captions, equivalent to the original representation. This seems contradictory, as we expected a loss in information content when compressing it to arbitrary 80-dimensions. This experiment indicates that the model is not explicitly utilizing the full expressiveness of the full 2048-dimensional representations. The model is able to learn from seemingly weak, structured information and can achieve performance that is close to that achieved using the full representation.



(a) Pool5

(b) Softmax

(c) Bag of objects

(d) Pseudo-random

Figure 1: Visualization of the t-SNE projection of initial representational space (left) vs. the transformed representational space (right). Please see `https://github.com/sheffieldnlp/whatIC` for original images.

4.2 Analyzing transformed image representations

Considering our earlier hypothesis as proposed in Section 3.5 whereby the conditioned RNN is learning some sort of a common 'visual-linguistic' semantic space, we explore the difference in representations in the initial representational space (I_m in Equation 1) and the transformed representational space (Im_{feat} in Equation 1). The transformation matrix W (Equation 1) is learned jointly as a subtask of the image captioning. We posit that image representations in the 256-dimensional transformed space will be more semantically coherent with respect to both images and captions. To visualize the two representational spaces, we use *Barnes-Hut t-SNE* [17] to compute a 2-dimensional embedding over the test split. In general, we found that images are initially clustered by visual similarity (*Pool5*) and semantic similarity (*Softmax*, *BOO*). After transformation, we observe that some linguistic information from the captions has produced different types of clusters. Figure 1 highlights some interesting observations regarding the changes in clustering across three different representations. For *Pool5*, images seem to be clustered by their visual appearance, for example snow scenes in Figure 1a, regardless of the subjects in the images (people or dogs). After transformation, separate clusters seem to form for snow scenes involving a single person, groups of people, and dogs. Interestingly, images of dogs in fields and snow scenes are also drawn closer together. *Softmax* (Figure 1b) shows many small, isolated clusters before transformation. After transformation, bigger clusters seem to be created – suggesting that the captions have again drawn related images together despite being different in the *Softmax* space. For *bag of objects* (Figure 1c), objects seem to be clustered by co-occurrence of object categories, for example toilets and kitchens are clustered since they share sinks. Toilets and kitchens seem to be further apart in the transformed space. We perform a similar analysis on the *pseudorandom* representations (Figure 1d). We observe that the initial representations have very little explicit information and do not cluster well, indicating that the pseudorandom vectors are indeed noisy. The projected representations, however, form clusters that mimic the projected space of the BOO cluster, demonstrating that the model is able to factorize the noisy representations in the visual-semantic projection space guided by information from the captions. Enlarged versions of the images in Figure 1 are also provided in the Appendix.

4.3 Domain dependency

We now demonstrate that end-to-end models are heavily reliant on datasets that have a similar training and test distribution. We posit that an IC system that performs similarity matching will not perform well on a slightly different domain for the same task. Demonstrating this will further validate our hypothesis that IC systems perform image matching to generate image captions.

We evaluate several models trained on MSCOCO on 1000 test image samples from the *Flickr30k* [34] dataset [6]. Like MSCOCO, Flickr30k is an image description dataset; however, unlike MSCOCO, the images have a different object distributions and the captions are slightly longer and more descriptive.

We evaluate the captions generated by our model with *ResNet152* pool5 representation and by two other state-of-the-art models pretrained on MSCOCO: (a) Self-Critical (SC) [22], based on self critical sequence training that uses reinforcement learning, and (b) Bottom Up and Top Down (TDBU) [1], based on top-down and bottom-up attention using object region proposals. Both state-of-the-art models are much more complex than the image-conditioned RNN language model. The results are summarized in Table 3.

We observe that the scores drop by a large margin. A similar observation was made by Vinyals et al. [28], and they alluded the drop in scores to the linguistic mismatch between the datasets. However, the out of training vocabulary words in the Flickr30k test set is only 8.6%. This suggests that there is more to the issue than a mere vocabulary mismatch. Typical sentences on Flickr30k are structurally different and generally longer, and the model is unable to generate good bigrams or even unigrams as is evident from B-1 and B-2 scores in Table 3.

5 Conclusions

We hypothesized that IC systems essentially exploit a *distributional similarity* space to 'generate' image captions by attempting to match a test image to similar training image(s) and generate an image caption from these similar images. Our study focused on the *image* side of image captioning: We varied the image representations while keeping the text generation component of an end-to-end CNN-RNN model constant. We found that regardless of the image representation, end-to-end IC systems seem to match images and generate captions in a visual-semantic subspace for IC. We conclude that:

[6] the test split is obtained from http://staff.fnwi.uva.nl/d.elliott/wmt16/splits.zip

(a) End-to-end IC models are remarkably capable of separating structure from noisy input representations, as demonstrated by **pseudo-random vectors**;

(b) End-to-end IC models suffer virtually no significant loss in performance when a high dimensional representation is **factorized** to a lower dimensional space;

(c) End-to-end IC models can **learn a joint visual-textual semantic subspace** by clustering images with similar visual and linguistic information together;

(d) End-to-end IC models rely on test sets having a **similar distribution** as the training set for generating good captions.

The observations above strengthen our distributional similarity hypothesis – that end-to-end IC models perform image matching and generate captions for a test image from similar image(s) from the training set – rather than performing actual image understanding. Our findings provide novel insights into what end-to-end IC systems are actually able to do, which previous work only suggests or hints at without concretely demonstrating. We believe our findings are important for the community to further advance work on image captioning in a more informed manner.

There is much scope for future work from the findings of this paper. One could examine the hidden states of the RNN model to better understand its behaviour and to further validate our distributional hypothesis. Understanding the theoretical formulation of the CNN-RNN architecture could also further help quantitatively confirm our hypothesis. Another useful direction would be to ascertain whether the distributional hypothesis also holds for more complex architectures, such as [1, 31]; our intuition is that the hypothesis would remain valid even for such models.

Acknowledgements. This work is supported by the MultiMT project (H2020 ERC Starting Grant No. 678017). The authors also thank the anonymous reviewers for their valuable feedback on an earlier draft of the paper.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the International Conference on Learning Representation (ICLR)*, 2016.

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[4] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015.

[5] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 100–105, 2015.

[6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 647–655, 2014.

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

[8] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1473–1482, 2015.

[9] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? Dataset and methods for multilingual image question answering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2296–2304, 2015.

[10] Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific computing*, 33(5):2580–2594, 2011.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 770–778, June 2016.

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

[13] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[14] Andrej Karpathy. *Connecting Images and Natural Language*. PhD thesis, Department of Computer Science, Stanford University, 2016.

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

[16] Remi Lebret, Pedro Pinheiro, and Ronan Collobert. Phrase-based image captioning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2085–2094, 2015.

[17] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9:2579–2605, 2008.

[18] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *Proceedings of the International Conference on Learning Representation (ICLR)*, 2015.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013.

[20] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 806–813, 2014.

[21] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[22] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representation (ICLR)*, 2015.

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112, 2014.

[26] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61(3):611–622, 1999.

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

[28] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, 2017.

[29] Josiah Wang, Pranava Swaroop Madhyastha, and Lucia Specia. Object counts! Bringing explicit detections back into image captioning. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2180–2193, 2018.

[30] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 203–212, 2016.

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 14, pages 77–81, 2015.

[32] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4904–4912, 2017.

[33] Xuwang Yin and Vicente Ordonez. Obj2Text: Generating visually descriptive language from object layouts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 177–187, 2017.

[34] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[35] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.