End-to-end Image Captioning Exploits Distributional Similarity in Multimodal Space [*]

Pranava Madhyastha **Josiah Wang** **Lucia Specia**
Department of Computer Science
University of Sheffield
{p.madhyastha,j.k.wang,l.specia}@sheffield.ac.uk

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, generally 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 (Young et al., 2014) and MSCOCO (Chen et al., 2015) are used to train a CNN-RNN based neural network IC system (Vinyals et al., 2017; Karpathy and Fei-Fei, 2015; Xu et al., 2015). 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) (Chen et al., 2015). However, in reality, the performance of end-to-end systems is still far from satisfactory according to metrics based on human judgement[2]. This task is thus currently far from being a solved problem.

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 induce that "*a boy is playing football*" 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 IC models have managed to infer that the phrase *football* 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 apparent strong performance of end-to-end systems is attributed to the fact that they exploit the *distributional similarity* in the multimodal feature space. To our best knowledge, our work is the first to provide empirical analysis of visual representations for the task of image captioning.

By 'distributional similarity' we mean that IC models essentially attempt to match 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 fact (Karpathy, 2016; Vinyals et al., 2017), but it has not been thoroughly investigated. This phenomenon could also be in part attributed to the fact that the datasets are repetitive and simplistic, with a virtually constant and predictable linguistic structure (Lebret et al., 2015; Devlin et al., 2015; Vinyals et al., 2017).

We empirically evaluate end-to-end IC systems where we vary the input image representation but keep the RNN text generation model constant. Our 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. We also analyze various types of image representations and their transformed versions.

[*] This is an abridged version of a recently published BMVC paper (Madhyastha et al., 2018)

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

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

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 381–383
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

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[3].

	Representation	B-4	M	C	S
	Random	0.07	0.11	0.07	0.03
Softmax	VGG19	0.19	0.20	0.61	0.13
	ResNet152	0.19	0.20	0.62	0.12
Penultimate	VGG19 (fc7)	0.22	0.21	0.69	0.14
	ResNet152 (pool5)	0.23	0.22	0.74	0.15
Embeddings	Top-k	0.19	0.20	0.63	0.13
BOO	Gold-Binary	0.22	0.22	0.75	0.15
	Gold-Counts	0.23	0.22	0.81	0.16
	YOLO-Coco	0.22	0.22	0.75	0.15
	YOLO-9k	0.21	0.20	0.68	0.13
Pseudo-random	Pseudorandom-Binary	0.21	0.21	0.73	0.14
	Pseudorandom-Counts	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 (BLEU-4), **M**eteor, **C**IDEr and **S**PICE scores.

We tabulate our observations from our experiments in Table 1 where we used standard end-to-end IC model (Vinyals et al., 2017) which is conditioned on the various image representations. We observe that utilizing standard bottleneck representations (penultimate) are slightly better than using the ImageNet class posteriors (softmax). However, we observe that better captions are obtained by using representations from explicit object detections.

We also introduce *pseudo-random vectors* which are derived from object-level representations as a control to evaluate IC systems. The pseudo-random representation is obtained using the object type information, but without actual object features. More specifically, $I_{pseudo} = \sum_{o \in \text{Objects}} f \times \phi_o$, where $\phi_o \in \mathcal{R}^d$ is an object-specific random vector and f is a scalar representing counts of the object category. Our results in Table 1 show that the models that utilize pseudo-random representations are able to perform competitively. The models in the current setup are remarkably capable of separating structure from noisy input. We further visualized the initial and projected representations in the setup and observed that while the initial pseudo-random representations were noisy, the projected ones closely resembled the bag-of-objects representations.

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.

We then perform experiments where IC models are conditioned on image representations *factorized and compressed to a lower dimensional space*. We experimented with three exploratory factor analysis based methods – Principal Component Analysis (PCA) (Halko et al., 2011), Probabilistic Principal Component Analysis (PPCA) (Tipping and Bishop, 1999) and Independent Component Analysis (ICA) (Hyvärinen et al., 2004). In all cases, we obtain 80-dimensional factorized representations on *ResNet152 pool5* ($2048D$) that is commonly used in IC. We summarize this experiment in Table 2. We observe that the representations obtained by all the factor 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 the information content when compressing it to arbitrary 80-dimensions. We observe 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 loss in performance, further strengthening our claim that IC models only perform similarity matching rather than image understanding. We conclude that the model is able to learn from a seemingly weak, structured information and is able to result in a performance that is close to that of a model that uses the full representation.

The observations above strengthen our distributional similarity hypothesis – that end-to-end IC performs image matching and generates 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 doing, which previous work only suggests or hints at.

[3] Our visualization and analysis can be found here: https://github.com/sheffieldnlp/whatIC

References

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

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

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

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

Andrej Karpathy. 2016. *Connecting Images and Natural Language*. Ph.D. thesis, Department of Computer Science, Stanford University.

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

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

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2018. End-to-end image captioning exploits distributional similarity in multimodal space. In *Proceedings of the British Machine Vision Conference (BMVC)*.

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

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

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. 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.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.