Learning Generative Image Manipulations from Language Instructions

Martin Längkvist, Andreas Persson, and Amy Loutfi

Center for Applied Autonomous Sensor Systems (AASS), Örebro University, Sweden
{martin.langkvist,andreas.persson,amy.loutfi}@oru.se

This paper studies whether a perceptual visual system can simulate human-like cognitive capabilities by training a computational model to predict the output of an action using language instruction. The aim is to ground action words such that an AI is able to generate an output image that outputs the *effect* of a certain *action* on an given *object*. Figure 1 illustrates the idea in principle where the input image contains several objects of different *shape*, *size*, and *color*, and an input instruction for how to manipulate one of the objects (i.e., *move*, *remove*, *add*, *replace*). The output of the model is a synthetic generated image that demonstrates the effect that the action has on the scene.[1]

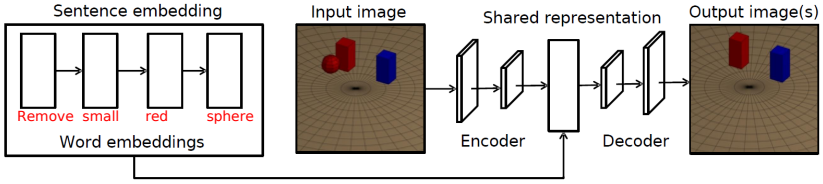

Fig. 1. A conceptual overview of the proposed model.

To visualize the *effect* of a certain *action*, a computational model must address a number of different sub-tasks, including; *1)* image encoding, *2)* language learning, *3)* relational learning, and *4)* image generation. In the literature, there have been works that combines some of these tasks for solving problems such as image captioning [8], image editing [2], image generation from text descriptions [5], visual question answering [6], paired robot action and linguistic translation [7], and Vision-and-Language Navigation (VLN)[1]. However, combining all the four sub-tasks, and how to learn their shared representations, is still an unaddressed challenge.

1 Proposed Model and Results

The proposed model, referred to as DCGAN+LSTM+RN, consists of a Deep Convolutional Generative Adversarial Networks (DCGAN)[4] as *image encoder* and *decoder*, a Long Short-Term Memory (LSTM) on a pre-defiend dictionary of 17 word representations as *language encoder*, and a Relational Network (RN)[6]

[1] This work is founded by the Swedish Research Council (Vetenskapsrådet), grant number: 2016-05321.

for *learning relations between object-pairs* and *merging image and language representations*. The model is trained in a Generative Adversarial Networks (GAN)[3] setting, which conditions both a source image and the action text description to generate a target image of the scene after the action has been performed. The model was trained on a dataset of 10000 generated image input-output pairs of 4 actions on objects of 3 types with 3 different colors and 2 sizes. The Root-Mean-Square-Error (RMSE) between generated and target images on a test set of 200 images and some visual results can be seen in Figure 2.A and Figure 2.B, respectively.

Using a language and relational model improves over the baseline. The purpose of the discriminator is to classify if the output image is *real with correct action* or either *fake with correct action* or *real but with wrong action*. Using a discriminator only slightly improves the results that make them look more realistic with less noise.

A. RMSE

Model	*Remove*	*Replace*	*Add*	*Move*	Overall
DCGAN (encoder only)	0.0407	0.0482	0.0441	0.0519	0.0457
DCGAN+LSTM+RN (encoder only)	0.0144	0.0222	0.0281	0.0264	0.0229
DCGAN+LSTM+RN (proposed)	0.0134	0.0208	0.0272	0.0249	0.0221

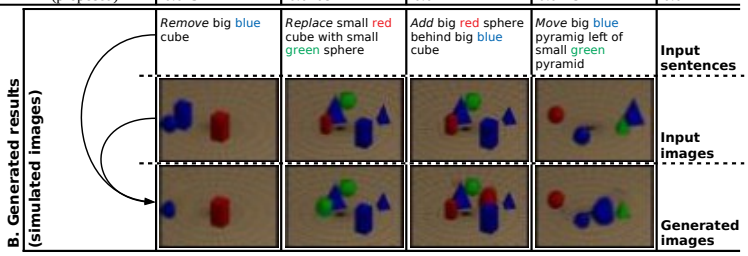

Fig. 2. A. RMSE between generated images and target images. **B.** Results on simulated images with four actions.

The pre-trained model was then tested on a sequence of five pre-processed real-world images, see Figure 3.A. The sequence of images was initially pre-processed, through a color-based segmentation approach (see Figure 3.B), before feed to the proposed model. Resulting generated images can be seen in Figure 3.D.

2 Conclusions and Future Work

This work combines an image encoder, language encoder, relational network, and image generator to ground action words, and then visualize the effect an action would have on a simulated scene. The focus in this work has been on learning meaningful shared image and text representations for relational learning and object manipulation. Directions of future work for adapting to real-world settings include: using pre-trained image and language encoders, and training on real-world images. Other directions for future work include generating sequences of images that illustrates *how* the action is performed and then performing the actions on a real robot.

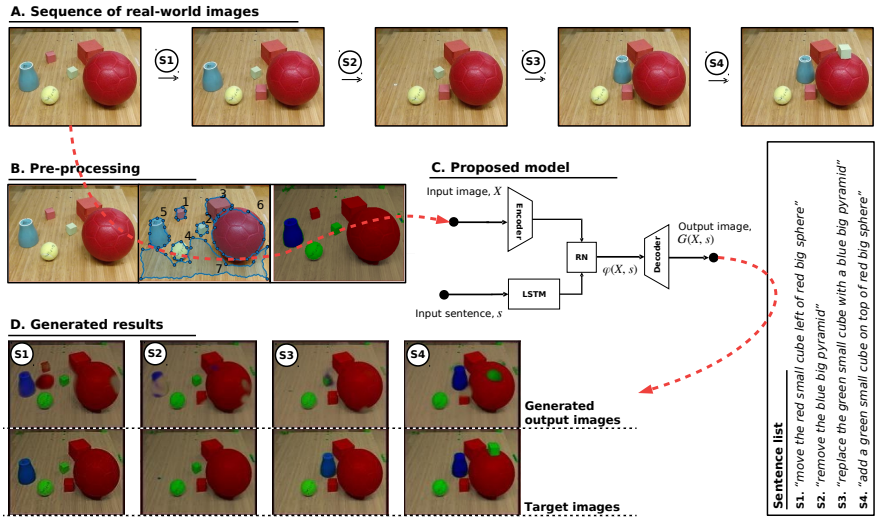

Fig. 3. Results on real-world images with pre-trained model.

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3674–3683 (2018)
2. Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X.: Language-based image editing with recurrent attentive models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8721–8729 (2018)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014), bluehttp://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
4. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
5. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. pp. 1060–1069. ICML'16, JMLR.org (2016), bluehttp://dl.acm.org/citation.cfm?id=3045390.3045503
6. Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.P.: A simple neural network module for relational reasoning. CoRR **abs/1706.01427** (2017), bluehttp://arxiv.org/abs/1706.01427
7. Yamada, T., Matsunaga, H., Ogata, T.: Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Robotics and Automation Letters **3**(4), 3441–3448 (2018)
8. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)