Supplementary Material

Perspective-Free Instructions

Although users naturally give instructions based on the perspective they know the recipient have, DSRs may still be required to understand perspective-free instructions. In the configuration addressed in Section 2, that is people with disabilities that would like the robot to bring everyday objects from somewhere in the house, it cannot be guaranteed that instructions have perspective-dependent expressions. Especially when considering the following cases:

- The target object may be located in a different room. Therefore, it is invisible to the user and the robot. This is a default setup of WRS (referred in Section 2), and is more practical than assuming that the user and robot can see the target object.
- The user is not able to know beforehand the robot's perspective when the target is in its view angle.
- The user may not be certain of the exact position of the target: unlike pieces of furniture, target objects are movable.

Another use of perspective-free system is emphasized in [19] where 3D object models are used to overcome the limitation of many image captioning method: the assumption of similar viewpoint in the training and testing phase. In our work, this limitation can be overcome by using M>1 images of the scene.

Introduction to the Attention Branch Network

Multi-ABN is inspired by the ABN [7]. The ABN is based on class activation mapping (CAM) networks [8, 13]. This line of attention research focuses on the production of image masks that, overlaid onto an image, highlight the most salient portions with respect to some given query or task. In essence, the CAM purpose is to identify salient regions used by a given class in an image classifier for visual explanation. The ABN builds visual attention maps from this approach. In the ABN, the CAM is extended to produce an attention mask for improving image classification. However, instead of directly using the attention network into the classifier, the ABN is decomposed into parallel branches to avoid deteriorating the classifier accuracy. The branches refer to the following sub-components of ABN:

1. an attention branch dedicated to producing the attention maps
2. a prediction branch predicting the likelihood of some label

Both the attention and prediction branches of ABN are classifiers. The attention maps are derived from the predicted label in the attention branch. Hence, this type of attention is built in a supervised manner. As an extension of CAM networks, ABN also allows visual explanation when extracting the attention maps. Such a feature is particularly desirable for qualitatively validating the network model.

Network parameters

The following table sums-up the parameters used for training the Multi-ABN.

Table 2: Parameter settings of the Multi-ABN

Multi-ABN Opt. method	Adam (Learning rate= $5e^{-4}$, $\beta_1 = 0.99$, $\beta_2 = 0.9$)
LSTM	3 layers, 1024-cell
MLP num. nodes	MLP_a: 1024, 1024
Vis. AB	Conv: 3x3x$\|V\|$, Att. Conv : 1x1x1
Ling. AB	Conv: 3x3x$\|V\|$, Att. Conv : 1x1x1, MLP: $\|V\|$
Batch size	32

Quantitative results

Table 3: Evaluation of Multi-ABN sentence generation for 5 trials. The average score and the standard deviation are reported. The Multi-ABN is compared with speaker model[22] using reinforcement learning as well as a baseline method using visual semantic embedding (VSE)[23], Multi-ABN with visual attention branch (VAB) only, and Multi-ABN with linguistic attention branch (LAB) only.

Method	Evaluation metric						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Speaker [22]	0.319 ± 0.029	0.201 ± 0.04	0.132± 0.04	0.102 ± 0.008	0.309 ± 0.015	**0.195** ± 0.017	0.802 ± 0.103
VSE	0.306 ± 0.013	0.199 ± 0.033	0.123 ± 0.012	0.073 ± 0.008	0.285 ± 0.033	0.108 ± 0.011	0.588 ± 0.083
Ours (VAB only)	0.323 ± 0.013	0.216 ± 0.010	0.143 ± 0.015	0.102 ± 0.010	0.333 ± 0.022	0.165 ± 0.031	0.824 ± 0.097
Ours (LAB only)	0.301 ± 0.020	0.250 ± 0.003	0.123 ± 0.020	0.099 ± 0.010	0.353 ± 0.051	0.142 ± 0.016	0.902 ± 0.107
Ours (Multi-ABN)	**0.390** ± **0.015**	**0.287** ± **0.009**	**0.184** ± **0.011**	**0.142** ± **0.012**	**0.359** ± **0.040**	0.193 ± 0.023	**1.048** ± **0.093**