Multimodal Attention Branch Network for Perspective-Free Sentence Generation

Aly Magassouba **Komei Sugiura**

Hisashi Kawai
National Institute of Information and Communications Technology
Japan
name.surname@nict.go.jp

Keywords: Domestic service robots, image captioning

Abstract: In this paper, we address the automatic sentence generation of fetching instructions for domestic service robots. Typical fetching commands such as "bring me the yellow toy from the upper part of the white shelf" includes referring expressions, *i.e.*, "from the white upper part of the white shelf". To solve this task, we propose a multimodal attention branch network (Multi-ABN) which generates natural sentences in an end-to-end manner. Multi-ABN uses multiple images of the same fixed scene to generate sentences that are not tied to a particular viewpoint. This approach combines a linguistic attention branch mechanism with several attention branch mechanisms. We evaluated our approach, which outperforms the state-of-the-art method on a standard metrics. Our method also allows us to visualize the alignment between the linguistic and visual features.

1 Introduction

The growth in the aged population has steadily increased the need for daily care and support. Robots that can physically assist people with disabilities [1] offer an alternative to overcoming the shortage of home care workers. This context has boosted the need for standardized domestic service robots (DSRs) that can provide necessary support functions as shown by [2, 3, 4].

Nonetheless, one of the main limitations of DSRs is their inability to naturally interact through language. Specifically, most DSRs do not allow users to instruct them with various expressions relating to an object for fetching tasks. By tackling this limitation, a user-friendly way to interact with DSRs could be provided to non-expert users.

Solving this task is particularly important for robots that perform manipulation tasks in home environments. Indeed, to understand ambiguous carry-and-place [5] or fetching [6] instructions in an end-to-end approach, a large number of samples of natural manipulation instructions are required. Currently, sophisticated DNNs are not fully utilized in robotics because most robotics data are small which easily lead to overfitting. Unfortunately, it is costly to obtain such data. Hence, methods to automatically augment or generate instructions data could drastically reduce the cost of building a large-scale dataset for DSRs by alleviating the burden of labelling from human experts.

In this light, our work addresses the task of automatic sentence generation for fetching instructions. This task consists of generating various natural fetching instructions given a target object in a image, *e.g.*,"*Go get me the empty bottle from the armchair on the right side* ." Natural sentences often contain referring expressions to designate a given target. However, generating referring expressions is challenging. Indeed, the many-to-many nature of mapping between the language and real world makes it difficult to generate such sentences.

In this paper, we propose the multimodal attention branch network (Multi-ABN) which is an extension of the attention branch network proposed in [7]. The initial attention branch network was proposed as an image classifier, inspired by class activation mapping (CAM) [8] structures, to infer

3rd Conference on Robot Learning (CoRL 2019), Osaka, Japan.

attention maps. It is composed of an attention branch that predicts an attention map and a perception branch that classifies images. This architecture is extended in Multi-ABN where several visual and linguistic attention branches are proposed to respectively infer the visual and linguistic attention maps. Indeed, instead of using a single image of a given scene, several snapshots of the same scene from different viewpoints are processed to generate perspective-free referring expressions. Our aim is to generate sentences that are not tied to the viewpoint of the human-annotated training image. From these attention maps, a long short-term memory (LSTM) network generates fetching instructions in the perception branch.

The main contributions of this paper are summarized as:

- We propose a Multi-ABN which generates fetching instructions based on multiple images from different perspectives of a fixed scene 4.
- Multi-ABN extends the existing methods by adopting visual and linguistic attention mechanisms based on class activation mapping structures.
- Multi-ABN outputs a visual explanation for the generated fetching instructions.

2 Related work

Building communicative robots that can understand ambiguous manipulation instructions generally require the fusion of multiple modalities, which are generally visual and linguistic. Several studies focus on understanding manipulation instructions in an end-to-end approach. For instance, [9] proposed a target object prediction method from natural language in a pick-and-place task environment, using a visual semantic embedding model. Similarly [10] tackled the same kind of problem using a two-stage model to predict the likely target from the language expression and the pairwise relationships between different target candidates. More recently, in a context related to DSRs, [6] proposed to use both the target and source candidates to predict the likely target in a supervised manner. In [6], the placing task was addressed through a GAN classifier network predicting the most likely destination from the initial instruction.

The proposed systems mainly focus on multimodal language grounding through referring expression comprehension. Complementary to these works, some recent studies have also focused on generating referring expressions to identify a target. In [11], the authors proposed several algorithms to generate referring expressions in a rule-based approach. In contrast, in [12], the authors used deep learning for estimating spatial relations to describe an object in a sentence. However, the set of spatial relationships is hand-crafted and known beforehand. We, instead, target an end-to-end approach that do not require hand-crafted or rule-based methods.

Multi-ABN is inspired by the attention branch network (ABN) [7]. The ABN is based on the CAM structure [8, 13] to build visual attention maps for image classification. In essence, a CAM is built to identify salient regions used by a given class in an image classifier. Attention mechanisms have also been used in different ways in image processing and natural language processing. In the context of image captioning, the authors of [14] proposed to generate image captions with hard and soft visual attention. This approach learns the alignment between the salient area of an image and the generated sequence of words. Multiple visual attention networks have also been proposed in [15] to solve visual question answering. However, most of these approaches use only a single modality for attention: visual attention. In contrast, we claim in this work that both linguistic and visual branch attention improve the sentence generation process.

To do so, we use annotated data obtained from the simulation environment SIGVerse [16]. Nowadays, many studies use simulated environments to collect synthetic data. Synthetic data tend to be increasingly photo-realistic and have the advantage of task repeatability as well as environment variation for a relatively low cost. Using such environments, various tasks such as grasping [17] or motion control [18] have been addressed.

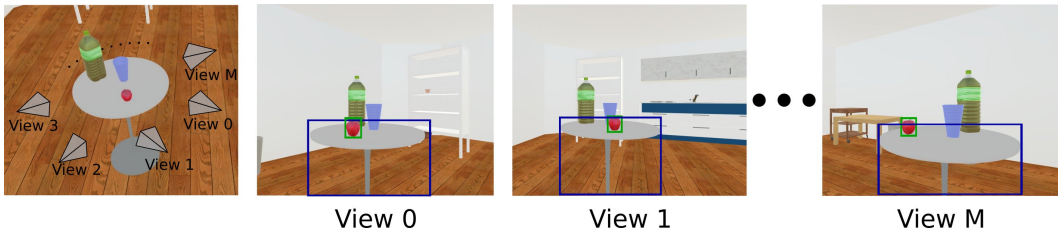


Figure 1: Source (blue) and target (green) samples of the WRS-VS dataset considering several perspectives. The perspective influences the validity of the instruction and the referring expressions that can be used. **Valid sentence** : "Bring me the apple that is near the glass on the kitchen table"/ **Invalid sentence**: " Bring me the apple on the left side of the blue glass".

3 Problem statement

3.1 Task Description

This study targets the generation of sentences for fetching instructions including referring expressions. Referring expressions usually describe an object using properties of the object with respect to landmark objects. A typical generated fetching instruction can be to " go get me the pink doll on the upper part of the shelf". In this instruction, the landmark is "the shelf" while "upper part" is a spatial referring expression. To generate such a sentence, our system assumes the following inputs and outputs:

- **Input**: a fixed scene observed from several perspectives.
- **Output**: the most likely generated sentence for a given target and source

The inputs of our system are more thoroughly described in Section 4. The terms *target* and *source* are defined as follows.

- **Target**: the daily life object (*e.g.*, apple or bottle) that the user intends for the robot to fetch.
- **Source**: the origin of the target, generally pieces of furniture such as shelves or drawers.

Unlike most methods proposed in the literature, our sentence generation method is based on several images of a fixed scene. Indeed, using single image to build a fetching instruction introduces a drawback when considering DSRs for manipulation task. This limitation is mainly related to DSRs that interact in a three-dimensional environment [19]. A DSR's view of a given scene is dynamic, *e.g.*, a target can be behind, on the left side, or on right side of the same landmark depending on the current robot pose (see Fig 1). Hence, to avoid generating referring expressions that are related to a given point of view of the scene, *e.g*, "the apple on the left side of the table", we use images from different perspectives of the same fixed scene. In this configuration, referring expressions such as "left of" or "right side of" are correct only if they are valid for all observations.

Several challenges should be tackled to generate valid fetching instructions. First, several objects may be of the same type as that of the target, so referring expressions should be used to disambiguate the target from the other objects. Second, several existing objects and sources may be used as landmarks for generating the referring expressions. However, the generated sentence should use referring expressions that do not imply any ambiguity of the target, independently of the point of view.

The standard evaluation metrics of our approach are based on the automatic metrics of image captioning that is BLEU, ROUGE, CIDEr and METEOR, as reported in the experimental section.

3.2 Task Environments

The sentence generation system should be general and flexible enough to be used for various scenarios. We therefore consider a simulated environment in which the task repeatability and various situations can be addressed at low cost. In this study, we use the simulated environments that were provided in the World Robot Summit 2018 Virtual Space (WRS-VS) challenge. The simulator is

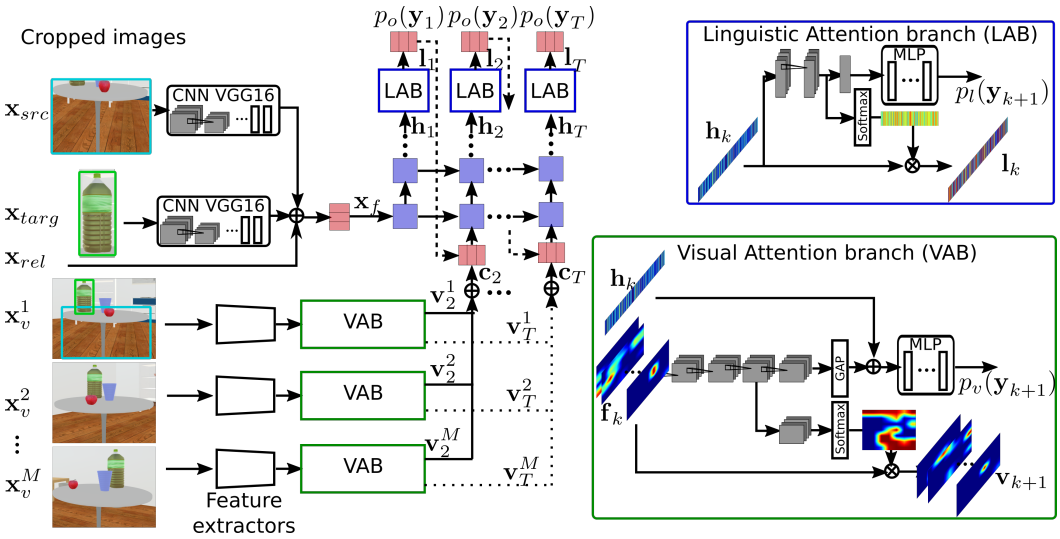


Figure 2: Proposed method framework: Multimodal attention branch mechanism that couples a linguistic attention branch to visual attention branches to generate fetching instructions

based on SIGVerse [16], which is a three-dimensional environment based on the Unity engine and is able to simulate interactions between agents and the environment. The WRS-VS consists of typical indoor environments as illustrated in Fig. 1, from which we built a dataset. In this environment, we use a DSR that records several snapshots of a given observable scene. From this context, our method should generate fetching instructions such as "*Give me the rabbit doll from the upper part of the shelf*".

4 Proposed method

Multi-ABN is composed of a linguistic attention branch as well as several visual attention branches for each different viewpoint. In the following we detail the input features, as well as the different branches (attention and perception branches) that have been used to address multimodality. The full network structure is given in Fig. 2. The aim of this network is to generate a sequence $Y = \{\mathbf{y}_1, \mathbf{y}_2 \dots \mathbf{y}_T\}$ where T is the length of a generated sentence and $\mathbf{y}_k \in \mathbf{R}^d$ for an embedding dimension d.

4.1 Input features

Let us consider $X = \{\mathbf{x}_n | n = 1, \dots, N\}$ a dataset composed of N samples. Hereinafter, for readability, we voluntarily omit the sample index n, so that $\mathbf{x}_n$ is written as $\mathbf{x}$ when further clarity is not required. Each sample $\mathbf{x}$ is characterized by the set of inputs:

$$\mathbf{x} = \{\mathbf{x}_v^1, \mathbf{x}_v^2, \dots, \mathbf{x}_v^M, \mathbf{x}_{src}, \mathbf{x}_{targ}, \mathbf{x}_{rel}\}. \tag{1}$$

The input $\mathbf{x}_v^j$ defines the set of image inputs taken from different viewpoints, where the superscript $j \in \{1, \dots, M\}$ defines the camera ID. Additionally, $\mathbf{x}_{src}$, $\mathbf{x}_{targ}$ and $\mathbf{x}_{rel}$ respectively denote the source image, target image and relation features of the target in the environment. It should be noted that the source, target and relation features are extracted only once from the main image $\mathbf{x}_v^1$ that is arbitrary chosen. As described in Section 4.3, these features condition the generated sentence.

Inputs $\mathbf{x}_{targ}$ and $\mathbf{x}_{src}$ are the cropped images of the target and source respectively. Inputs $\mathbf{x}_{rel}$ denotes the target-source, target-image and source-image spatial relation. Each of these relations are characterized by the following:

$$\mathbf{r}_{m/n} = \left[\frac{x_m}{W_n}, \quad \frac{y_m}{H_n}, \quad \frac{w_m}{W_n}, \quad \frac{h_m}{H_n}, \quad \frac{w_m h_n}{W_m H_n}\right] \tag{2}$$

where (x_m, y_m, w_m, h_m) are the horizontal position, vertical position, width and height of the component m, while W_n and H_n are the width and height of the component n. As a result, the relation features are defined as $\mathbf{x}_{rel} = \{\mathbf{r}_{targ/_{src}}, \mathbf{r}_{targ/_v}, \mathbf{r}_{src/_v}\}$ with a dimension $d_{rel} = 15$.

4.2 Attention branches

4.2.1 Visual attention branches

The attention branch [7] allows us to build both the linguistic and visual attention map, based on the extracted feature maps. We consider two different types of feature maps, linguistic and visual detailed below.

Multi-ABN is composed of M visual attention branches, that is for the images $\mathbf{x}_v^j$. Each of these visual attention branch takes as input visual feature maps denoted $\mathbf{f}_k$. These feature maps are obtained from a convolutional feature extraction of $\mathbf{x}_v^j$. In this paper, we based our feature extractor on VGG16[20]. Note that other feature extractor such as ResNet[21] could also be used. $\mathbf{f}_k$ then corresponds to the output from the 5^{th} convolutional block of the VGG network. Each feature map $\mathbf{f}_k$ has a dimension 14x14x512. A visual attention branch outputs visual feature maps $\mathbf{v}_{k+1}$ weighted by a visual attention mask. To do so, inspired by the CAM structure [8], the visual feature maps are encoded through four convolutional layers. These convolutional layers are followed by a global average pooling (GAP) and a two-layer multilayer perceptron (MLP) denoted as MLP_a. Prior to the first layer of MLP_a, the visual features are concatenated with the linguistic feature map $\mathbf{h}_k$ (see next section). The likelihood $p_v(\mathbf{y}_{k+1})$ is then predicted. In parallel, a visual attention map $\mathbf{a}_k$ is created by an additional convolution and sigmoid normalization of the third convolutional layer of the visual attention branch. This attention map allows to selectively focus on certain parts of an image related to the predicted sequence. The output visual feature maps are then obtained by a masking process given by:

$$\mathbf{v}_k = \mathbf{a}_k \odot \mathbf{f}_k, \tag{3}$$

where $\odot$ denotes the Hadamard product.

4.2.2 Linguistic attention branch

In addition a linguistic attention branch takes as input the linguistic feature maps $\mathbf{h}_k$. $\mathbf{h}_k$ is simply defined as the output (or hidden state) of an LSTM generating the instruction sequence Y. This LSTM network is detailed in the next section. The linguistic feature map $\mathbf{h}_k$ is encoded through 1-dimensional convolution layers followed by a single fully connected layer so as to output likelihood $p_l(\mathbf{y}_{k+1})$. Linguistic attention map $\mathbf{a}_l$ is obtained from the second convolutional layer that is convoluted in an additional layer and normalized by a sigmoid activation function. This attention map allows to selectively focus on a area of the LSTM state the also encodes all the previous states. Similarly to visual attention branches, the output $\mathbf{l}_k$ of linguistic attention branch is given by:

$$\mathbf{l}_k = \mathbf{a}_l \odot \mathbf{h}_k \tag{4}$$

4.3 Perception branch

The perception branch is a classifier that predicts the likelihood of $p(\mathbf{y}_{k+1})$ in a sequence of length T. The perception branch takes as input the concatenation of all weighted visual feature maps $\mathbf{v}_k^j$ that is referred as $\mathbf{c}_k$ and the weighted linguistic feature map $\mathbf{l}_k$, as well as the target, source and relation features. The architecture of the perception branch is based on a multilayer LSTM network. The perception also outputs linguistic feature map $\mathbf{h}_k$ that is simply the hidden state of each LSTM cell. Note that because each embedded word $\mathbf{y}_k$ is predicted sequentially, the last hidden state also corresponds to the output of the LSTM. More thoroughly, the LSTM is initialized by the latent space feature $\mathbf{x}_f$ obtained by embedding and concatenating the target $\mathbf{x}_{targ}$, the source $\mathbf{x}_{src}$ and the relation feature $\mathbf{x}_{rel}$. In a compact formulation, the first hidden state can be written as

$$\mathbf{h}_1 = \text{LSTM}(\mathbf{x}_f). \tag{5}$$

It should be mentioned that the forget, memory, output and hidden state variables are voluntarily omitted for more concision. In the following steps, considering an iteration k, with $k > 0$, each hidden state is defined as follows

$$\mathbf{h}_k = \text{LSTM}(E(\mathbf{c}_k \oplus \mathbf{y}_{k-1})), \tag{6}$$

where $\oplus$ indicates a concatenation operation and $E(\cdot)$ is an embedding function. In this configuration, $\mathbf{c}_k$ can be considered as the visual context of the current LSTM state. Eventually, to predict the likelihood $p_o(\mathbf{y}_{k+1})$ in the sequence, the weighted linguistic feature map $\mathbf{l}_k$ is processed in a embedding layer.

4.4 Loss functions

The global training loss function of the network is the sum of the attention branch loss L_{att} and the perception branch loss L_{per} so that

$$L = L_{att} + L_{per}. \tag{7}$$

Perception loss L_{per} is defined as a cross-entropy function in which the class of $\mathbf{y}_{k+1}$ is predicted through

$$L_{per} = -\sum_{n}\sum_{m} y^*_{nm} \log p(y_{nm}), \tag{8}$$

where y^*_{nm} denotes the label given to the m-th dimension of the n-th sample, and y_{nm} denotes its prediction. The attention loss L_{att} depends on the visual attention loss and linguistic attention loss, which are also both cross-entropy loss functions, as defined in Eq. (8), that enable to build the corresponding attention maps.

5 Experiments

5.1 Dataset

We evaluated our method with the WRS-VS dataset introduced in Section 3 and illustrated in Fig. 1. We collected $308 \times$ M images. In the following experiment, we set M $= 3$, which means that there were three different images from each given scene. We annotated 1015 targets with 2015 sentences in the training set and 34 targets with 74 different sentences in the validation set. The annotation was performed by an expert user and was intended to be perspective-free. For a target in a scene, the annotator was given M images from different perspectives, and was instructed to give a sentence that would be valid for all M images. This data set has an average of 3.4 targets per image, and 9.5 words for each instruction. The vocabulary set V is composed of 233 unique words.

5.2 Experimental Setup

The parameter settings of Multi-ABN are summarized in the supplementary material. We describe first the different attention branches that compose the Multi-ABN. Each visual attention branch (noted Vis. AB) uses 2D convolutional layers of size $3 \times 3 \times \|V\|$ before the global average pooling layer. To generate each visual attention map, a convolutional layer of dimension $1 \times 1 \times 1$ is used. Because we consider M $= 3$ different images, a two-layer MLP$_a$ is used to encode the different weighted visual feature maps concatenated with the linguistic feature map. In parallel, the linguistic attention branch (noted Ling. AB) uses 1D convolutional layers of size $3 \times 3 \times \|\text{V}\|$ followed by a single-layer embedding of dimension $\|\text{V}\|$. Similarly to the visual case, the linguistic attention map is obtained by processing the features with a convolutional layer of dimension $1 \times 1 \times 1$.

In the perception branch, the LSTM has $N = 3$ layers, with each cell having dimension $d = 1,024$. The network is trained with an Adam optimizer with a learning rate of $5e^{-4}$, considering a batch size of 32 samples.

5.3 Quantitative results

As mentioned in Section 3, we use standard image captioning metrics to evaluate the performance of Multi-ABN. For the sake of exhaustiveness we report the results of baseline metrics BLEU score (1-gram to 4-gram) as well as more evolve scoring systems ROUGE, CIDEr and METEOR. These scores are reported in each column of Table 3. To avoid bias in the scoring system, each generated sentence was run against multiple corresponding annotated sentences.

The Multi-ABN was compared with a baseline method [22], in which a speaker model is used to generate sentences. For fair comparison, because the speaker model is only adapted to $M = 1$

Table 1: Evaluation of Multi-ABN sentence generation. 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	0.201	0.132	0.102	0.309	**0.195**	0.802
VSE	0.306	0.199	0.123	0.073	0.285	0.108	0.588
Ours (VAB only)	0.323	0.216	0.143	0.102	0.333	0.165	0.824
Ours (LAB only)	0.301	0.250	0.123	0.099	0.353	0.142	0.902
Ours (Multi-ABN)	**0.390**	**0.287**	**0.184**	**0.142**	**0.359**	0.193	**1.048**

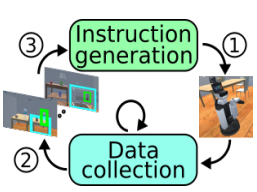
images, we concatenated $M = 3$ images into a single one to form the input of the speaker model. The same method was applied for comparison to another baseline that is the visual semantic embedding architecture [23]. As reported in Table 3, Multi-ABN outperforms the speaker model under ROUGE, CIDEr and BLEU score, while METEOR evaluation does not make emphasize any significant difference. In comparison to VSE method, Multi-ABN performs significantly better for all metrics, in particular, CIDEr score is improved by 0.46 points. These results are also supported by a statistical analysis in the supplementary file.

In addition, several ablation tests were conducted to isolate and emphasize the contribution of each attention branch mechanism. We compare our results, Multi-ABN with visual attention branch only (VAB) and Multi-ABN with linguistic attention branch (LAB). The results in Table 3 show that the Multi-ABN drastically improved all the metrics. The visual and linguistic attention branches each improved the baseline visual semantic architecture. The LAB improved the sentence generation quality more in terms of ROUGE and CIDEr, while the BLEU and METEOR scores were better with the VAB only. This suggests that LAB leads to a better generalization and variety in the produced sentence, while the VAB is better able to mimic the dataset sentences.

5.4 Qualitative results

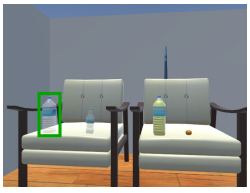
5.4.1 Sentence generation

In the following, because of limited space, we illustrate our results with only a single image (*i.e.*, $\mathbf{x}_v^1$), however two auxiliary images were used to train Multi-ABN and generate the fetching instruction. Qualitative results of our method are illustrated in Fig. 3. Multi-ABN can be applied in a framework (see subfigure (a)) where the generated sentence are instructed to a robot ①. In a second step while completing the fetching task, the robot collects additional data ② that can be used to train Multi-ABN in return ③. Subfigures (b) and (c) present correct fetching instructions, while subfigure (d) show an erroneous sentence generation. Indeed in the latter, the referring expression does not allow to disambiguate the target from the other large bottle.



(a)

(b) I want a rabbit item on the upper part of shelf



(c) I want an empty large plastic bottle placed on the left hand sofa

(d) Can you get a large plastic bottle on the side table

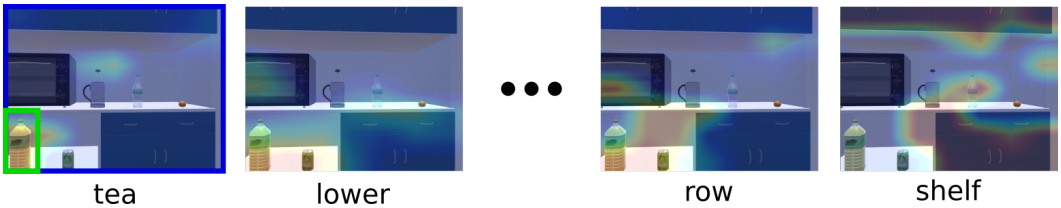
Figure 3: Generated sentences by our method. Solid green rectangles represent the target. Only the first image $\mathbf{x}_v^1$ of each sample are given. Subfigures (b) and (c) show correct predictions while subfigure (d) shows an erroneous generation

5.4.2 Visualization of attention maps

Similarly to methods based on visual attention, Multi-ABN can also exhibit the visual alignment between text and image. These alignments are depicted in Fig. 4, where the attention map for each generated word is depicted on the image $\mathbf{x}_v^1$.

(a) Bring me the small item on the right-sided armchair.

(b) Pick up the yellow toy from the white shelf.



(c) Take the tea on the lower row of the shelf.

Figure 4: Multi-ABN visual attention evolution for different sentence steps of sentence generation for the most representative words. The visual attention is updated to the relevant parts of the image.

These results confirm the relevance of the visual attention for generating a sequence of words. Multi-ABN is able to learn the correspondence between linguistic and visual features since the network was able to focus on relevant areas for each word.

6 Conclusion

Motivated by the development of communicative DSRs, we developed the multimodal attention branch network, Multi-ABN, that generates natural fetching instructions including referring expressions. We summarize the following important contributions of the paper:

- Multi-ABN is an attention branch network that generates sentences based on visual and linguistic attention. Multi-ABN yields better under BLEU, ROUGE and CIDEr metrics than a baseline method such as [22].
- Multi-ABN outperforms visual only or linguistic only attention branch networks, which emphasizes the contribution of both linguistic and visual modalities.
- Multi-ABN is able to generate perspective-free fetching instructions by the use of several visual attention branches related to different viewpoints of the same scene.

Nonetheless, Multi-ABN is not limited to sentence generation and could be used in various robotic researches such as multimodal language understanding or object recognition with specific inputs and network structure. In future work, we plan to extend our work with a physical experimental study with non-expert users. Additionally, we plan to apply Multi-ABN on a fully communicative DSR scheme that couples sentence generation and natural language comprehension for fetching tasks.

Acknowledgments

This work was partially supported by JST CREST and SCOPE.

References

[1] S. W. Brose, D. J. Weber, et al. The role of assistive robotics in the lives of persons with disability. *American Journal of Physical Medicine & Rehabilitation*, 89(6):509–521, 2010.

[2] L. Piyathilaka and S. Kodagoda. Human Activity Recognition for Domestic Robots. In *Field and Service Robotics*, pages 395–408, 2015.

[3] C.-A. Smarr, T. L. Mitzner, J. M. Beer, A. Prakash, T. L. Chen, C. C. Kemp, and W. A. Rogers. Domestic Robots for Older Adults: Attitudes, Preferences, and Potential. *International Journal of Social Robotics*, 6(2):229–247, 2014.

[4] L. Iocchi, D. Holz, J. Ruiz-del Solar, K. Sugiura, and T. Van Der Zant. RoboCup@ Home: Analysis and Results of Evolving Competitions for Domestic and Service Robots. *Artificial Intelligence*, 229:258–281, 2015.

[5] A. Magassouba, K. Sugiura, and H. Kawai. A Multimodal Classifier Generative Adversarial Network for Carry and Place Tasks From Ambiguous Language Instructions. *IEEE RA-L*, 3 (4):3113–3120, Oct 2018.

[6] A. Magassouba, K. Sugiura, A. Trinh Quoc, and H. Kawai. Understanding natural language instructions for fetching daily objects using gan-based multimodal target-source classification. *arXiv preprint arXiv:1906.06830*, 2019.

[7] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition*, 2016.

[9] J. Hatori et al. Interactively Picking Real-World Objects with Unconstrained Spoken Lnguage Instructions. In *IEEE ICRA*, pages 3774–3781, 2018.

[10] M. Shridhar and D. Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *RSS*, 2018.

[11] L. Kunze, T. Williams, N. Hawes, and M. Scheutz. Spatial referring expression generation for hri: Algorithms and evaluation framework. In *2017 AAAI Fall Symposium Series*, 2017.

[12] F. I. Doğan, S. Kalkan, and I. Leite. Learning to generate unambiguous spatial referring expressions for real-world environments. *arXiv preprint arXiv:1904.07165*, 2019.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[15] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] T. Inamura, J. T. C. Tan, K. Sugiura, T. Nagai, and H. Okada. Development of robocup@ home simulation towards long-term large scale hri. In *Robot Soccer World Cup*, pages 672–680. Springer, 2013.

[17] K. Bousmalis et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *Proc. IEEE ICRA*, pages 4243–4250, 2018.

[18] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

[19] V. Cohen, B. Burchfiel, T. Nguyen, N. Gopalan, S. Tellex, and G. Konidaris. Grounding language attributes to objects using bayesian eigenobjects. *arXiv preprint arXiv:1905.13153*, 2019.

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

[21] K. He, S. Zhang, X.and Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE CVPR*, pages 770–778, 2016.

[22] L. Yu, H. Tan, M. Bansal, and T. L. Berg. A joint Speaker Listener-Reinforcer Model for Referring Expressions. In *CVPR*, volume 2, 2017.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.