IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 13, 2020, accepted December 20, 2020, date of publication December 23, 2020,
date of current version January 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046719

Cross-Lingual Visual Grounding

WENJIAN DONG[1], MAYU OTANI[2], NOA GARCIA[3], YUTA NAKASHIMA[3], (Member, IEEE), AND CHENHUI CHU[4]

[1]École Polytechnique, 91128 Palaiseau, France
[2]CyberAgent, Inc., Shibuya 150-0042, Japan
[3]Institute for Datability Science, Osaka University, Suita 565-0871, Japan
[4]Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

Corresponding author: Chenhui Chu (chu@i.kyoto-u.ac.jp)

This work was supported in part by the Japan Science and Technology Agency (JST) ACT-I, in part by the Microsoft Research Asia Collaborative Research Grant, and in part by the Japan Society for the Promotion of Science (JSPS) through the Grant-in-Aid for Young Scientists under Grant 19K20343,.

ABSTRACT Visual grounding is a vision and language understanding task aiming at locating a region in an image according to a specific query phrase. However, most previous studies only address this task for the English language. Although there are previous cross-lingual vision and language studies, they work on image and video captioning, and visual question answering. In this paper, we present the first work on cross-lingual visual grounding to expand the task to different languages to study an effective yet efficient way for visual grounding on other languages. We construct a visual grounding dataset for French via crowdsourcing. Our dataset consists of 14k, 3k, and 3k query phrases with their corresponding image regions for 5k, 1k, and 1k training, validation and test images, respectively. In addition, we propose a cross-lingual visual grounding approach that transfers the knowledge from a learnt English model to a French model. Despite that the size of our French dataset is 1/6 of the English dataset, experiments indicate that our model achieves an accuracy of 65.17%, which is comparable to the accuracy 69.04% of the English model. Our dataset and codes are available at `https://github.com/ids-cv/Multi-Lingual-Visual-Grounding`.

INDEX TERMS Visual grounding, cross-lingual, vision and language.

I. INTRODUCTION

Studies on various vision and language tasks, such as image captioning [1] and visual question answering [2], have significantly promoted the research on joint vision and language understanding. Visual grounding, which aims at finding a specific region in an image corresponding to a query phrase, plays a fundamental role in enhancing the performance of many joint vision and language tasks. Since the emergence of the first work of visual grounding in [3], research efforts have been dedicated to improve its accuracy [4]–[11].

These studies are targeted mostly on English because large-scale visual grounding datasets are only available in English. Although English is the major language, visual grounding could be also important for other languages for joint vision and language understanding in those languages. Studies on English, however, may not necessarily address the visual grounding task for other languages, as the query phrases may be completely different for different languages, and thus a visual grounding model is strongly tied to the specific language used for training. Constructing dedicated datasets in other languages in a similar scale to the English ones could be a solution, but its cost is expensive. Therefore, visual grounding on languages other than English remains an unaddressed problem. The objective of our work is to study an effective yet efficient way for visual grounding on other languages.

To this end, in this paper, we present the first work on cross-lingual visual grounding that transfers the knowledge obtained from an English visual grounding model to another language. Cross-lingual transfer works because the same visual concept is shared despite different languages. Cross-lingual studies have been already applied to visual captioning [12] and visual question answering [13] but not yet to visual grounding. A high-performance cross-lingual visual grounding model not only provides efficient query localization for other languages but also has the potential to improve the performance of multimodal multilingual tasks, such as multimodal machine translation [14].

We exemplify our idea with the French language. We construct a French visual grounding dataset via crowdsourcing, which consists of 14k, 3k, and 3k query phrases with their

The associate editor coordinating the review of this manuscript and approving it for publication was Sara Dadras.

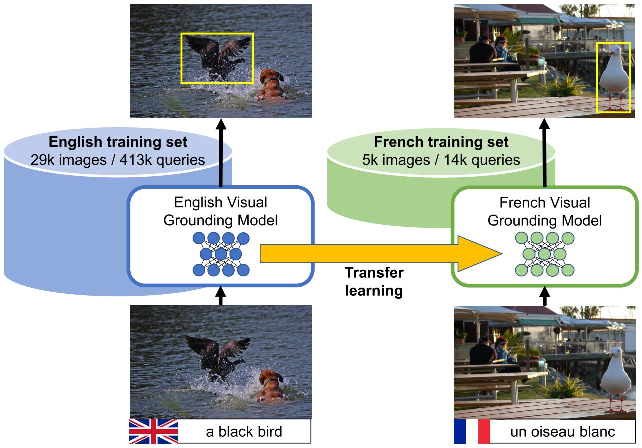


FIGURE 1. **Overview of cross-lingual visual grounding.**

corresponding image regions for 5k, 1k, and 1k training, validation and test images, respectively. As our French dataset is small in size compared to the English dataset (442k query phrases in 31k images), we adopt the state-of-the-art English model [11] and apply transfer learning techniques [15] to fine-tune the model for French (see Figure 1). Our experimental results show that our model achieves a comparable accuracy to the English model on the test split of our French dataset. Our contribution is two-fold:

- We introduce the first visual grounding dataset in a non-English language for cross-lingual visual grounding.
- We propose a transfer learning approach for visual grounding and experimentally verify its effectiveness.

II. RELATED WORK

A. ENGLISH VISUAL GROUNDING

Visual grounding is a task to find an image region that corresponds to a given phrase in a caption. Visual grounding and object detection often share the same computer vision techniques for proposal generation, but the former surpasses the latter in terms of the versatility of query it could handle. Object detection only handles pre-defined classes (for example, the PASCAL-VOC dataset [16] has 20 categories and the Microsoft COCO (MS COCO) dataset [17] has 80 categories), while visual grounding has no pre-defined categories and has the capacity to handle an unlimited number of categories in principle. Moreover, visual grounding also handles nouns with modifiers while object detection does not.

Many visual grounding studies have been conducted for English. Plummer *et al.* [3] released the Flickr30k entities dataset and proposed a method based on canonical correlation analysis (CCA) [18] that learns joint embeddings of phrases and image regions. Wang *et al.* [4] proposed a two-branch neural network for joint phrasal and visual embeddings. Fukui *et al.* [5] used multimodal compact bilinear pooling to fuse phrasal and visual embeddings. Rohrbach *et al.* [6] proposed a method to first detect a candidate region for a given phrase and then reconstruct the phrase using the detected region. Wang *et al.* [7] proposed an agreement-based method, which encourages semantic relations among phrases to agree with visual relations among regions. Yeh *et al.* [8] proposed a framework that can search over all possible regions instead of a fixed number of region proposals. Plummer *et al.* [9] used spatial relationships between pairs of phrases connected by verbs or prepositions. Chen *et al.* [10] proposed a reinforcement learning-based model that rewards the grounding results with image-level context. Yu *et al.* [11] improved the region proposal network by training it on the Visual Genome dataset [19] to increase the diversity of object classes and attribute labels, which achieved the state-of-the-art performance. In this paper, we apply the model of [11] for cross-lingual visual grounding.

Inspired by the success of pre-training language models such as BERT [20], vision and language pre-training on large image caption datasets such as the conceptual captions dataset [21] has been promoted such as [22]–[24]. Those vision and language pre-training models differ from the model architecture, but we refrain the details here because this is beyond the focus of this work. Vision and language pre-training is evaluated on tasks including visual grounding. However, same to previous studies, the visual grounding task is still limited to English [22]–[24].

B. CROSS-LINGUAL VISION AND LANGUAGE

As image/video captioning and visual question answering are the most representative vision and language tasks, we introduce cross-lingual work for these two tasks here.

1) IMAGE AND VIDEO CAPTIONING

Although there are some large-scale image captioning datasets, such as the MS COCO [25] and Flickr30k [26], many of them only provide English captions when they were first released. These datasets have been extended by adding captions in other languages for cross-lingual study. Miyazaki and Shimizu [12] created the YJ Captions, which is a Japanese image captioning dataset using a part of the images from MS COCO, and applied transfer learning for the task. Yoshikawa *et al.* [27] further enlarged the collection of Japanese captions for MS COCO and released the STAIR Captions dataset. There are also some extensions for Chinese, such as [28], [29]. Li *et al.* [28] presented comparison of Chinese caption datasets constructed by crowdsourcing and machine translation. Li *et al.* [29] added Chinese captions and tags for MS COCO. For video captions, Chen and Dolan [30] collected short video clips and captions in many different languages. They recruited monolingual speakers to make a large-scale and linguistically diverse dataset.

Nakayama *et al.* [31] translated the English captions in the Flickr30k entities dataset into Japanese. They also linked the translated Japanese entities to their corresponding regions. They applied a CCA based method similar to [3] to conduct visual grounding experiments on the created Japanese dataset. Different from their work, our visual grounding dataset is created for French. Moreover, we study transfer

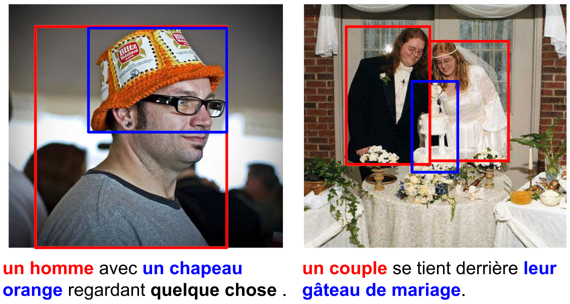

un homme avec un chapeau orange regardant quelque chose .
un couple se tient derrière leur gâteau de mariage.

FIGURE 2. Two examples of French captions with corresponding image regions in our dataset.

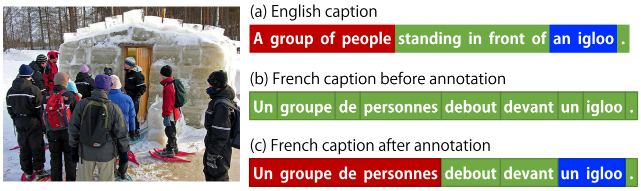

(a) English caption
A group of people standing in front of an igloo .
(b) French caption before annotation
Un groupe de personnes debout devant un igloo .
(c) French caption after annotation
Un groupe de personnes debout devant un igloo .

FIGURE 3. An example of our annotation task. An English sentence (a) and its corresponding French caption (b) are presented to a worker at the initial state of our task, where entities in (a) are in different colors. In the French caption after annotation (c), corresponding phrases are colored.

learning to improve French visual grounding by transferring knowledge from the English model.

2) VISUAL QUESTION ANSWERING

Visual question answering (VQA) is a task to generate a natural language answer to a question about the image content. Antol *et al.* [32] firstly published the VQA dataset. This task has also been extended in different languages for cross-lingual study. Gao *et al.* [13] published a Chinese VQA dataset, which includes an English translation of the annotation, and Shimizu *et al.* [33] created a Japanese VQA dataset.

III. FRENCH VISUAL GROUNDING DATASET

In this section, we introduce the approach we use to construct our French visual grounding dataset, named as *Flickr5k French entities*. Figure 2 shows two examples of French captions in our dataset with corresponding image regions.

A. RELATED DATASETS

Visual grounding requires (image, phrase, image region) triplets for training. We build our dataset in French based on the benchmark dataset Flickr30k and its derivatives:

Flickr30k [26] is composed of 31k images and 158k captions in English (5 captions per image).

Flickr30k entities [3] enhances the Flickr30k dataset by identifying 443k entities in the Flickr30k's captions with their corresponding image region.

Multi30k [14] extends Flickr30k into French, German, and Czech. For French in particular, the Multi30k dataset translates 1 of 5 English captions per image in the original Flickr30k dataset.

B. ANNOTATION PROCESS

Based on the data already available in those datasets, we construct the *Flickr5k French entities* dataset in the following way: we use the images from the Flickr30k dataset, entities and image regions in the Flick30k entities dataset, and French translations from the Multi30k dataset. We then identify entities in French captions that correspond to those in the Flickr30k entities dataset via crowdsourcing. As the English entities have been linked to their corresponding image regions in the Flickr30k entities dataset, the identified French entities can be also linked with the same regions. We follow the train-validation-test split used in the Multi30k dataset (the 5k training sentences that we annotated in this work are the first 5k sentences in the training set of the Multi30k dataset). The statistics of the annotated captions is shown in the first three columns in Table 1.

We construct our dataset in this way instead of creating a new dataset from scratch. This is advantageous in two ways: 1) As numerous studies have been conducted on those benchmark datasets, it will make comparison among different works easier; 2) It can contribute to multimodal multilingual tasks such as multimodal machine translation on the Multi30k dataset.

We use Amazon Mechanical Turk[1](AMT) as our crowdsourcing platform and develop our own interface dedicated to the task.[2] We present image, English caption, and French caption to workers[3] (see Figure 3). Each entity phrase in English caption is shown in a different color. The workers are asked first to click on an entity of interest in the English caption and then click on the corresponding French words (multiple words involved in most cases). Non-entity words remain in the green color. Figure 3 shows an example of our annotation task and Table 1 shows the statistics of our dataset in detail.

C. QUALITY CONTROL

In order to facilitate quality control, we put 12 questions in each HIT (Human Intelligence Task, the basic work unit on AMT). This also makes the annotation speed of workers faster

[1] https://www.mturk.com/

[2] We did an experiment using the GIZA++ toolkit to automatically align the words between English and French based on 31k sentence pairs. We then compared the auto-aligned sentences against the human-annotated sentences based on 1k sentences (the first 1k sentences in the training set), and found that 5% of words are incorrectly assigned. For more distant language pairs such as English-Japanese, it has been shown that word alignment accuracy is significantly lower than that of English-French, and thus word alignment cannot be directly used for our purpose in general. As we wanted to guarantee the quality of our dataset, we chose to use human annotation instead of word alignment. Annotating by correcting the errors made by the word aligner can be another strategy that might be more efficient, but we had the concern that this is more difficult to control the quality of annotation.

[3] We limited workers to English or French speaking countries (France, Belgium, Canada, Switzerland, Australia, USA, UK), and we required "master qualification" to work for us.

TABLE 1. The statistics of our French visual grounding dataset. Numbers in bracket are the corresponding statistics of the original Flickr30k entities dataset in English. Note that when we count vocabularies in this table, we do not conduct stemming, just convert all text into lower case.

Data split	Images involved	Caption(s) per image	Total entities	Entities with image region	Vocabulary size	Vocabulary size with freq ≥ 3
Train	5,000 (29,000)	1 (5)	16,187 (507,718)	13,834 (413,627)	3,110 (14,382)	1,059 (6,709)
Val	1,014 (1,014)	1 (5)	3,319 (17,956)	2,756 (14,526)	1,266 (2,885)	309 (1,108)
Test	1,000 (1,000)	1 (5)	3,311 (17,577)	2,789 (14,476)	1,229 (2,758)	330 (1,026)

thanks to the reduction of the time for submitting responses and loading new questions.[4]

Among 12 questions per HIT, we include 2 quality control questions in each HIT. Quality control questions are randomly chosen from a pool of accepted answers on previously annotated questions. We randomly shuffle normal questions and quality control questions within a HIT. We evaluate the quality of the HIT by computing the word-wise accuracy u over the two quality control questions:

$$u = \frac{1}{n}\sum_{i}\frac{T_i}{T_i + F_i}$$

where T_i and F_i are the numbers of words annotated correctly and incorrectly, respectively, on the i-th quality control question of the HIT ($i \in \{1, 2\}$), and n is the number of quality control questions in one HIT being set to 2.

D. ANNOTATION ERROR ANALYSIS

We reviewed the submitted HITs with u being less than 0.9 (about 4.2% of all HITs), and we found that most of the discrepancies on quality control questions which led to low u score are upon ambiguous translations, and the annotation quality on 10 normal questions was generally good. Moreover, we randomly reviewed more HITs. In total, we manually checked roughly 15% of the entire annotated data.

Annotation errors in our dataset mainly fall into the following three categories:[5]

1) **Errors in content words** are due to negligence of workers. We estimated that the errors in this category contaminated 0.5% of all annotated phrases.
2) **Errors in functional words** have limited influence in the meanings of phrases due to the nature of functional words. Most errors in this category are inclusion of unnecessary prepositions or missing of articles. These errors contaminated about 2% of all annotated phrases.
3) **Errors in Flick30k entities or Multi30k** also result in errors in our annotation as we built our dataset on top of them. This category is estimated to occupy about 1% of all annotated phrases.

IV. VISUAL GROUNDING MODEL

This section introduces the model for visual grounding and our transfer learning strategy to train the model for other languages with much smaller volume of training data.

A. OUR MODEL

Inspired by Yu *et al.* [11], we implement a simple and powerful visual grounding model shown in Figure 4,[6] which has the capability of efficiently picking up the best region proposal among a large number of region proposals.

The input image and query phrase are first processed separately. For the image, we use the Faster-RCNN [34] to get N region proposals, each of which is composed of a visual feature vector $f_v \in \mathbb{R}^{d_v}$ and a spatial feature vector $f_s \in \mathbb{R}^5$. As in [35], f_s consists of the normalized coordinates of the top-left and bottom-right corners as well as the area of the region proposal. More specifically,

$$f_s = \left[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{wh}{WH}\right]$$

where (x_{tl}, y_{tl}) and (x_{br}, y_{br}) are the coordinates of the top-left and bottom-right corners of the proposal; $w = x_{br} - x_{tl}$ and $h = y_{br} - y_{tl}$; W and H are the width and height of the image. Figure 5 shows proposals generated by Faster R-CNN upon an example image.

For the query phrase, we build a vocabulary index table based on our training set with an adding token for unknown words, which replaces a word not in the training set. Because our training set is relatively small and different word forms like singular or plural are informative for the task, we only transform all words into lowercase but do not do stemming. We transform the query phrase into a index sequence q, which is then fed into a word embedding layer and a single-layer long short-term memory (LSTM) unit. Hidden states of LSTM form a continuous-space representation for a word given all preceding words. We use the hidden state $f_q \in \mathbb{R}^{d_q}$ of the last word as the query phrase representation:

$$f_q = \text{LSTM}(\text{embed}(q))$$

where embed$(\cdot)$ converts indices in q into word embeddings and LSTM$(\cdot)$ gives the last word's hidden state.

We then concatenate the visual features f_v and f_s for each region proposal and the textual features f_q, i.e.,

$$f = \text{concat}(f_v, f_s, f_q)$$

[4] A screenshot of our user interface can be found in the supplementary material.

[5] The error rate for each category is estimated based on our manually checked HITs. It is calculated at phrase level, i.e. if there is one word misannotated in a phrase, this phrase will be counted as a failure. For each category, some examples of errors are listed in the supplementary material.

[6] We do not adopt the regression loss in [11], because it decreases the accuracy in our preliminary experiments.

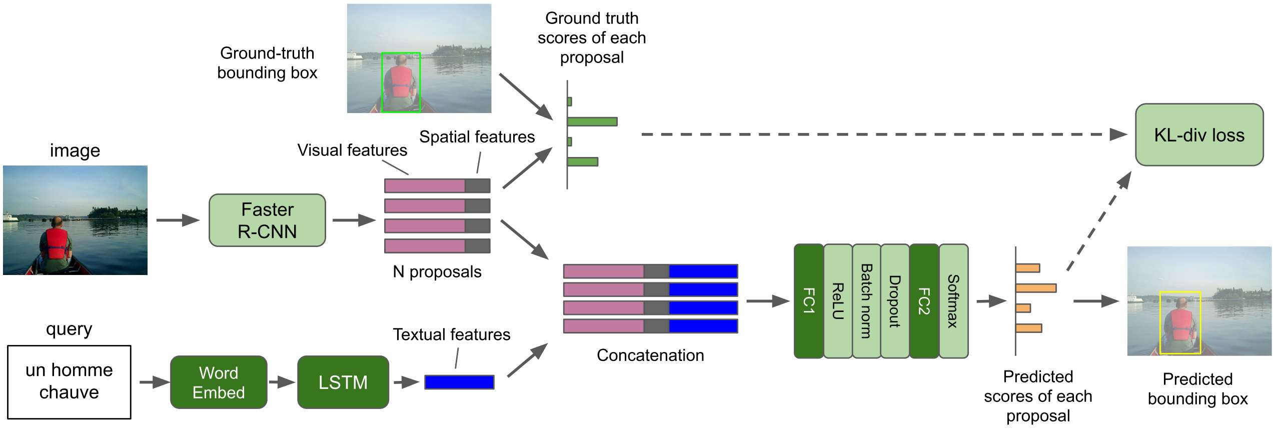


FIGURE 4. The workflow of our visual grounding model. Note that the Faster R-CNN module is standalone and we do not optimize its parameters. The query "un homme chauve" means "a bald man." The modules in deep green (Word Embed, LSTM, FC1 and FC2) are the modules whose parameters will be transferred from the English model to the French model.

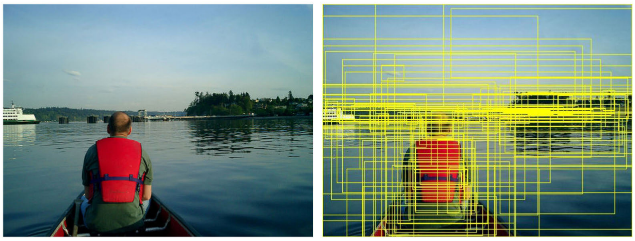


FIGURE 5. An example of an image (left) and region proposals for the image generated by Faster R-CNN (right).

and put all proposals in a matrix F:

$$F = [f_1, \ldots, f_N].$$

where $F \in \mathbb{R}^{N \times d_0}$ with $d_0 = d_v + 5 + d_q$.

Next, to rank the proposals according to the query phrase, we pass F through a fully-connected layer (FC1), a batch normalization layer, a dropout layer, and a second fully-connected layer (FC2) consecutively. Formally,

$$F_1 = \mathrm{ReLU}(FW_1 + b_1\mathbf{1}^T)$$

where $W_1 \in \mathbb{R}^{d_0 \times d_1}$, $b_1 \in \mathbb{R}^{d_1}$, $\mathbf{1}$ is a vector with all N elements being 1. Then

$$\begin{aligned} F_1' &= \mathrm{dropout}(\mathrm{batchnorm}(F_1)) \\ S &= \mathrm{softmax}(F_1' W_2 + b_2\mathbf{1}^T) \end{aligned}$$

where $W_2 \in \mathbb{R}^{d_1}$, $b_2 \in \mathbb{R}$, and $S = [s_1, \ldots, s_N] \in [0, 1]^N$ is the scores for N proposals.

The ground truth score g_n of the n-th proposal is the IoU score between the proposal and the ground truth image region with thresholding. Let r_n and t denote the n-th proposal and the ground truth image region for the query phrase. We compute g_n by

$$g_n' = \begin{cases} \mathrm{IoU}(r_n, t) & \text{if } \mathrm{IoU}(r_n, t) > \eta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where η is the threshold. We normalize the score over the N proposals by

$$g_n = \frac{g_n'}{\sum_i g_i'}.$$

For training our model, we use the Kullback-Leibler divergence as loss L:

$$L = \sum_n g_n \log \frac{g_n}{s_n}$$

B. TRANSFER LEARNING

As in Figure 4, our model consists of vision and language branches, which are merged to predict the score for each proposal. The vision branch can be independent from the language branch; therefore, the vision branch trained for a certain language can be reusable for another language. On the other hand, the language branch needs to be re-trained for different languages. This can be seen as an analogy of our own human experience. For an English speaker who already knows "computer," when he learns the French word "ordinateur," he will not need to learn this word as a French person did in 1950s who had never seen any computer in his life. Instead he will just map this French word to the concept of "computer" already in his mind. Such a transfer learning strategy greatly facilitates acquisition of foreign languages for human beings. We do the same thing for our machine learning model for visual grounding.

To implement such an idea for transferring knowledge from English to French visual grounding, we first train our English model from scratch with the Flickr30k entities dataset. Then we initialize a French model in the same network architecture, except the vocabulary size in the word embedding layer. The other parameters of the French model are initialized with those of the English model. Figure 6 illustrates the workflow of this transfer learning strategy. For the word embedding layer, we test five different strategies:

Random mapping assigns a word vector of the English model to each French word randomly.

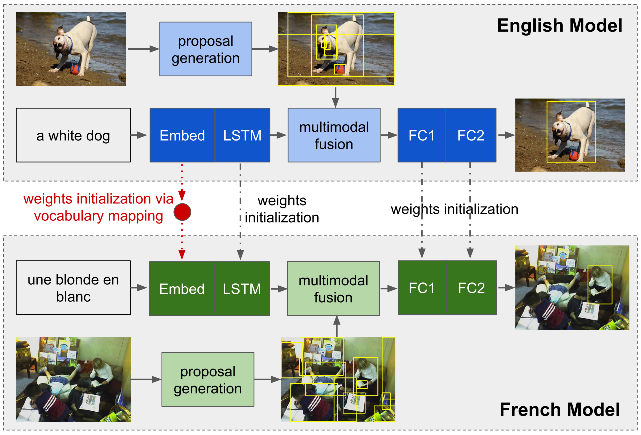


FIGURE 6. The workflow of our transfer learning model. The word embedding layer is treated with special strategies, while other layers in the French model are transferred directly from the English model. The Embed, LSTM, FC1, and FC2 modules are fine-tuned.

Frequency-based mapping counts the word frequency in the English and French training sets separately and sorts them in a descending order. The English word vector is assigned to the French word at the same position in the French's sorted word list as the English one.

General dictionary mapping (GDM) uses a general purpose dictionary, e.g., Google Translate, to map each French word to an English word. If an English word suggested for a certain French word by the general-purpose dictionary does not used in our English dataset, we randomly assign an English word vector to this French word. When the dictionary suggests a multi-word English phrase for one French word, we use the last word in the English phrase.

Dataset-specific dictionary mapping (DSDM) is the same as the GDM approach except that the dictionary is built from the dataset. We use the GIZA++ toolkit[7] to align words in the training set of the Multi30k English-French datasets and then calculate the lexical translation probability of French and English word pairs. We choose the English word with the highest translation probability as the translation of a French word.

The general-purpose dictionary mapping approach has wide applicability in many real scenarios because most languages, even low-resource ones, have dictionaries between them and English. This provides our approach great portability to address visual grounding problems even for many low-resource languages.

V. EXPERIMENTS

A. SETTINGS

For training the English model, we generally followed the configuration suggested in [11]. We used the Faster-RCNN with ResNet101 provided by Anderson *et al.* [36] to get $N = 100$ region proposals for each image. The dimension d_v of visual features was set to 2,048. The dimension of word embedding was set to 300. For the LSTM layer's output, we used $d_q = 1,024$. For FC1's output, we used $d_1 = 512$. For FC2's output, we set $d_2 = N = 100$. When calculating the Kullback-Leibler divergence, we added $\epsilon = 10^{-7}$ to all s_n in order to avoid zero division. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was 0.001 with an exponential decay rate 0.7. The dropout rate was 0.5, and the mini-batch size was 64. We trained the model by 10 epochs, saving the model after each epoch. Finally we picked up the model with the best performance on the validation set as the chosen model, and evaluated it on the test set.

TABLE 2. Comparison between our English model and existing ones.

Model	Accuracy
PL-CLC [9]	55.85
QRC-Net [10]	65.14
DDPN [11]	**73.30**
Our model	69.04

Following previous studies, we treated a prediction as successful when the predicted image region overlaps the ground truth with IoU > 0.5, and we calculated the percentage of such successful predictions within a certain test set as the accuracy over this test set. The threshold η in Equation (1) was set to 0.5, the same value as the criterion between a successful and an unsuccessful prediction.

B. RESULTS

1) PERFORMANCE OF OUR ENGLISH MODEL

We first report the performance of our English model in Table 2.[8] We can see that, although our model performs slightly worse compared to [11], it is comparable to the state-of-the-art.

2) CROSS-LINGUAL GROUNDING PERFORMANCE

For our *Flickr5k French entities* dataset, we compared our transfer learning model with three baselines.

- Training from scratch: all layers are trained on the French data from scratch.
- GDM w/o TL: the word embedding layer is initialized by GDM, but all other layers are trained from scratch.
- Ground after Translate: translates the French test set into English word by word using our general dictionary, and then uses the English model to ground.

To better understand the performance of different transfer learning strategies on different volumes of training data, we tested them with 1k and 3k sets. The 5k set is the entire training set of *Flickr5k French entities* dataset, while the 1k and the 3k take their corresponding images from the beginning of the training set.

Table 3 summarizes the performances. We can see that despite the small volume of the training set, our French model reaches comparable accuracy to the English model. As for

[7] http://code.google.com/p/giza-pp

[8] Recent pre-training vision and language representation studies also tested on English visual grounding such as [22]–[24]. However, all these studies did not report results on the Flicker30k entities dataset, but on another dataset of RefCOCO+. Therefore, we did not add them into Table 2.

TABLE 3. The accuracy scores for the three baselines and our model with different transfer learning strategies (split by the middle line). "1k", "3k", and "5k" mean training on entities for 1k, 3k, and 5k images, respectively. Note that "Ground after Translate" is not trained on our Flickr5k French entities dataset.

Strategy	1k	3k	5k
Training from scratch	52.67	58.66	60.93
GDM w/o TL	53.75	59.48	60.13
Ground after Translate		62.28	
Random mapping	51.74	58.09	60.69
Frequency mapping	49.52	56.83	59.51
GDM	60.92	63.79	65.17
DSDM	**62.39**	**66.08**	**66.39**

TABLE 4. The performance of our model with freezing certain layers when training the French model. "T" refers to "trained" and "F" refers to "fixed." We use the GDM strategy and train our model on 5k images.

Embed	LSTM	FC1	FC2	Accuracy
F	F	F	F	60.40
T	F	F	F	62.63
T	T	F	F	**65.44**
T	T	T	F	65.26
T	T	T	T	65.17

word embedding initialization, random and frequency mapping perform worse than training from scratch. When using a dictionary to map the word embeddings, the performance is significantly improved.[9] GDM performs significantly better than GDM w/o TL especially when the training data is smaller, indicating the importance of transfer learning. 5k performs better than 3k, and 3k performs significantly better than 1k, indicating that our model can be further improved by accumulating the French training data. Surprisingly, Ground after Translate shows a high accuracy, but it is still worse than GDM with 3k or 5k images. DSDM outperforms GDM, showing the ability of our model when parallel captions are available for generating a dictionary.

3) EFFECTS WHEN FREEZING CERTAIN LAYERS

For all results in Table 3, the parameters of all layers are trained in the French model, except for the vision branch. We also investigated the performance of our model when we freeze some layers during training. Table 4 summarizes the results. In the table, we see that the best result is achieved when we only fine-tune the embedding and the LSTM layer while fixed the two fully connected layers. These experimental results confirm our intuition in Section IV-B, that the superficial difference in languages can be handled in the lower layers (the layers close to text) and the conceptual representations (the layers close to image in our case) for different languages can be shared.

4) APPLICABILITY TO OTHER LANGUAGES

As different languages have different word orders, we also conducted experiments to investigate the influence of word orders to the performance of our transfer learning-based approach. We trained a model with randomly shuffling the word order in each French query. The results are shown in Table 5. We can see that our model performs similarly to the original and shuffled word orders. This indicates that word orders should not matter when we apply our model to different languages.[10]

TABLE 5. Accuracy when word order is shuffle. We used the GDM strategy and all layers are trained over entities for 5k images.

	Accuracy
Original word order	65.17
Shuffled word order	64.91

C. DISCUSSION

We analyzed the French visual grounding results to investigate the strengths and weaknesses of our model.

1) SUCCESSFUL EXAMPLES

The top row of Figure 7 shows some successful examples. We found that our model has the capability to localize objects described by a wide variety of words; it can also localize objects in different amounts (e.g., 1, 3, and 5 in Figure 7) or in different sizes (e.g., 2 and 4 in Figure 7). Our model also works for both fully and partially presented objects (e.g., 2 and 4 in Figure 7).

2) UNSUCCESSFUL EXAMPLES

The bottom row of Figure 7 shows some unsuccessful examples. We found that some are with resembling rivals (e.g., 6 and 7 in Figure 7); some are with different amounts (e.g., 8 and 9 in Figure 7); some are due to the lack of context information (e.g., 10 in Figure 7); others are failures due to unknown words. Note that our English model also makes similar mistakes for these examples. These problems point out future research directions. For example, we may need to take language and visual context into account to handle some queries; we may also conceive a more advanced approach to deal with words not appeared in training set by sub-words, or an external knowledge base.

Though comparable, our French model still performs slightly worse than our English model. Therefore, we also investigated examples where our French model fails but our English model succeeds to analyze the weakness of transfer learning. Figure 8 shows such examples. We found that in many cases our French model tends to focus on a part of the object while our English model can ground correctly (e.g., 1 vs 6, 2 vs 7, 3 vs 8 in Figure 8); there are also cases where our French model encompasses redundant regions. However, we did not observe meaningful trends. We think that these cases happen due to the small size of the French dataset and more training data could help.

To summarize, we can see that our French model performs comparable to the English model. However, due to

[9] We provide examples of word mapping in the supplementary material.

[10] Note that complicated queries such as "the dog holding a brown toy," where changing the word order of nouns will change the semantic meaning do not exist in our dataset. "the dog holding a brown toy" is split into two queries of "the dog" and "a brown toy" in Flick30k entities and thus also split in our dataset.

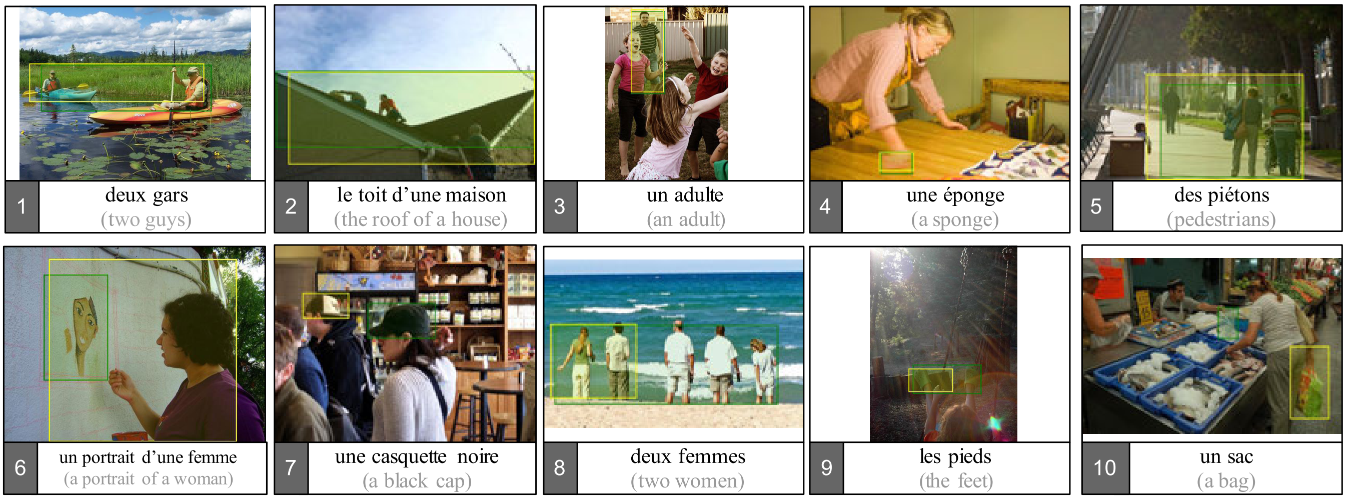


FIGURE 7. Examples of successful (top) and unsuccessful (bottom) visual grounding. The green boxes are the ground truth image region, and the yellow boxes are the prediction by our French model.

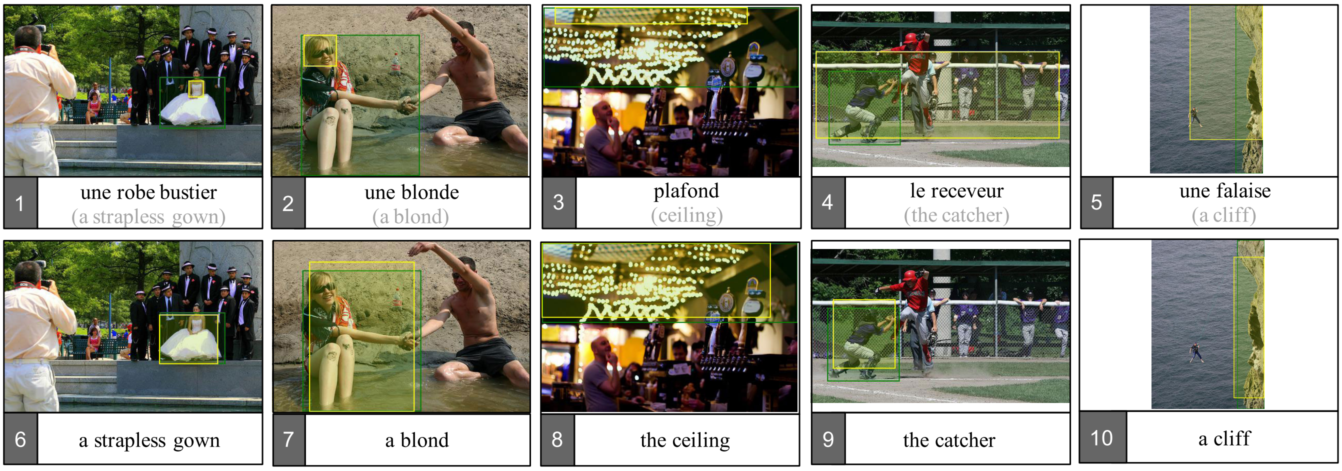


FIGURE 8. Examples of visual grounding examples that our French model fails (top), but our English model succeeds (bottom). The green boxes are the ground truth image region, and the yellow boxes are the prediction.

the limitations of the original English model in lacking of context and unknown words, our model still fails in these cases. We believe recent advances in sub-word based context-aware pre-training visual and language representations such as [22]–[24] are more robust to these issues and plan to study cross-lingual visual grounding based on these models.

VI. CONCLUSION

In this paper, we presented the first work on cross-lingual visual grounding to study an effective yet efficient way for visual grounding on other languages. We constructed a visual grounding dataset in French and proposed to transfer the knowledge from a state-of-the-art English visual grounding model to the French one. Experimental results showed that our transfer learning-based approach can achieve an accuracy comparable to the English model, even with a small French dataset.

As future work, we plan to conduct experiments on languages distant from English, such as Chinese and Japanese, and verify the effectiveness of our approach on these languages. This work still needs to annotate a small visual grounding dataset in another language, making it hard to extend to other languages. We plan to study robust cross-lingual representations so that we can get rid of the expensive annotation process in this work.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[2] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. V. D. Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Understand.*, vol. 163, pp. 21–40, Oct. 2017, doi: 10.1016/j.cviu.2017.05.001.

[3] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer Image-to-Sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.

[4] L. Wang, Y. Li, and S. Lazebnik, ''Learning deep structure-preserving image-text embeddings,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.

[5] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, ''Multimodal compact bilinear pooling for visual question answering and visual grounding,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2016, pp. 457–468.

[6] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, ''Grounding of textual phrases in images by reconstruction,'' in *Proc. ECCV*, Oct. 2016, pp. 817–834.

[7] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng, ''Structured matching for phrase localization,'' in *Proc. ECCV*, Oct. 2016, pp. 696–711.

[8] R. Yeh, J. Xiong, W. W. Hwu, M. Do, and A. G. Schwing, ''Interpretable and globally optimal prediction for textual grounding using image concepts,'' in *Proc. NIPS*, 2017, pp. 1909–1919.

[9] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, ''Phrase localization and visual relationship detection with comprehensive image-language cues,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1928–1937.

[10] K. Chen, R. Kovvuri, and R. Nevatia, ''Query-guided regression network with context policy for phrase grounding,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 824–832.

[11] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, and D. Tao, ''Rethinking diversified and discriminative proposal generation for visual grounding,'' in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1114–1120, doi: 10.24963/ijcai.2018/155.

[12] T. Miyazaki and N. Shimizu, ''Cross-lingual image caption generation,'' in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, vol. 1, Aug. 2016, pp. 1780–1790.

[13] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, ''Are you talking to a machine? Dataset and methods for multilingual image question answering,'' in *Proc. NIPS*, 2015, pp. 2296–2304.

[14] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, ''Findings of the second shared task on multimodal machine translation and multilingual image description,'' in *Proc. 2nd Conf. Mach. Transl.*, 2017, pp. 215–233.

[15] S. J. Pan and Q. Yang, ''A survey on transfer learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ''The Pascal visual object classes (VOC) challenge,'' *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ''Microsoft coco: Common objects in context,'' in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[18] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, ''Canonical correlation analysis: An overview with application to learning methods,'' *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li, ''Visual genome: Connecting language and vision using crowdsourced dense image annotations,'' 2016, *arXiv:1602.07332*. [Online]. Available: http://arxiv.org/abs/1602.07332

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. ACL Conf. NAACL HLT*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[21] P. Sharma, N. Ding, S. Goodman, and R. Soricut, ''Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, Jul. 2018, pp. 2556–2565.

[22] J. Lu, D. Batra, D. Parikh, and S. Lee, ''Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,'' in *Proc. NeurIPS*, 2019, pp. 13–23.

[23] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, ''Vl-bert: Pre-training of generic visual-linguistic representations,'' in *Proc. ICLR*, 2020, pp. 2–16.

[24] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, ''Uniter: Universal image-text representation learning,'' in *Proc. ECCV*, Aug. 2020, pp. 104–120.

[25] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, ''Microsoft COCO captions: Data collection and evaluation server,'' 2015, *arXiv:1504.00325*. [Online]. Available: https://arxiv.org/abs/1504.00325

[26] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, ''From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,'' *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.

[27] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, ''STAIR captions: Constructing a large-scale japanese image caption dataset,'' in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, Jul. 2017, pp. 417–421.

[28] X. Li, W. Lan, J. Dong, and H. Liu, ''Adding chinese captions to images,'' in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 271–275.

[29] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu, ''COCO-CN for cross-lingual image tagging, captioning, and retrieval,'' *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2347–2360, Sep. 2019.

[30] D. L. Chen and W. B. Dolan, ''Building a persistent workforce on mechanical turk for multilingual data collection,'' in *Proc. HCOMP*, 2011, p. 6.

[31] H. Nakayama, A. Tamura, and T. Ninomiya, ''A visually-grounded parallel corpus with phrase-to-region linking,'' in *Proc. LREC*, May 2020, pp. 4204–4210.

[32] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, ''VQA: Visual question answering,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[33] N. Shimizu, N. Rong, and T. Miyazaki, ''Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps,'' in *Proc. CICLing*, Aug. 2018, pp. 1918–1928.

[34] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[35] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, ''Modeling context in referring expressions,'' 2016, *arXiv:1608.00272*. [Online]. Available: http://arxiv.org/abs/1608.00272

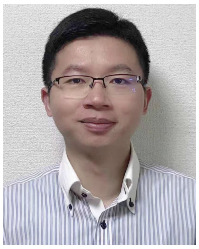
[36] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, ''Bottom-up and top-down attention for image captioning and visual question answering,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

WENJIAN DONG received the B.S. degree in physics from Wuhan University, in 2016. He is currently pursuing the double Engineering degrees in data science at the École Polytechnique and Télécom Paris.

MAYU OTANI received the B.S. degree from Kyoto University, in 2013, and the M.S. and Ph.D. degrees in engineering from the Nara Institute of Science and Technology, in 2015 and 2018, respectively. She is currently a Research Scientist with CyberAgent, Inc. Her research interests include video understanding and multimodal machine learning.

NOA GARCIA received the B.S. degree in telecommunication engineering from the Universitat Politècnica de Catalunya, Spain, in 2012, and the Ph.D. degree in computer science from Aston University, U.K., in 2019. She is currently a Researcher with Osaka University, Japan. Her research interests include applications on high-level visual understanding at the intersection of computer vision, natural language processing, and machine learning.

YUTA NAKASHIMA (Member, IEEE) received the B.E. and M.E. degrees in communication engineering and the Ph.D. degree in engineering from Osaka University, Osaka, Japan, in 2006, 2008, and 2012, respectively. From 2012 to 2016, he was an Assistant Professor with the Nara Institute of Science and Technology. He is currently an Associate Professor with the Institute for Datability Science, Osaka University. His research interests include computer vision and machine learning and their applications. His main research interest includes video content analysis using machine learning approaches. He is a member of ACM, IEICE, and IPSJ.



CHENHUI CHU received the B.S. degree in software engineering from Chongqing University, in 2008, and the M.S. and Ph.D. degrees in informatics from Kyoto University, in 2012 and 2015, respectively. He is currently a Program-Specific Associate Professor with Kyoto University. His research interests include natural language processing, particularly machine translation, and multimodal machine learning.