

Received October 29, 2021, accepted November 22, 2021, date of publication November 30, 2021, date of current version December 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3131393

Imageability- and Length-Controllable Image Captioning

MARC A. KASTNER¹, KAZUKI UMEMURA², ICHIRO IDE^{3,2}, (Senior Member, IEEE), YASUTOMO KAWANISHI^{4,2}, (Member, IEEE), TAKATSUGU HIRAYAMA^{2,5}, (Member, IEEE), KEISUKE DOMAN⁶, (Member, IEEE), DAISUKE DEGUCHI², (Member, IEEE), HIROSHI MURASE², (Life Fellow, IEEE), AND SHIN'ICHI SATOH¹, (Member, IEEE)

¹Digital Content and Media Sciences Research Division, National Institute of Informatics, Chiyoda-ku, Tokyo 101-8430, Japan

²Graduate School of Informatics, Nagoya University, Chikusa-ku, Nagoya, Aichi 464-8601, Japan

³Mathematical and Data Science Center, Nagoya University, Chikusa-ku, Nagoya, Aichi 464-8601, Japan

⁴Guardian Robot Project, Information Research and Development and Strategy Headquarters, RIKEN, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

⁵Faculty of Human Environment, University of Human Environments, Okazaki, Aichi 444-3505, Japan

⁶School of Engineering, Chukyo University, Toyota, Aichi 470-0393, Japan

Corresponding author: Marc A. Kastner (mkastner@nii.ac.jp)

This work was supported in part by the JSPS KAKENHI 16H02846 Program, in part by the Microsoft CORE-16 Research Program, and in part by a joint-research project between the National Institute of Informatics and Nagoya University.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Image captioning can show great performance for generating captions for general purposes, but it remains difficult to adjust the generated captions for different applications. In this paper, we propose an image captioning method which can generate both imageability- and length-controllable captions. The imageability parameter adjusts the level of visual descriptiveness of the caption, making it either more abstract or more concrete. In contrast, the length parameter only adjusts the length of the caption while keeping the visual descriptiveness on a similar degree. Based on a transformer architecture, our model is trained using an augmented dataset with diversified captions across different degrees of descriptiveness. The resulting model can control both imageability and length, making it possible to tailor output towards various applications. Experiments show that we can maintain a captioning performance similar to comparison methods, while being able to control the visual descriptiveness and the length of the generated captions. A subjective evaluation with human participants also shows a significant correlation of the target imageability in terms of human expectations. Thus, we confirmed that the proposed method provides a promising step towards tailoring image captions closer to certain applications.

INDEX TERMS Machine learning, semantics, task analysis, image captioning, psycholinguistics.

I. INTRODUCTION

Image captioning shows great performance in generating captions for general purposes and receives great attention in the research community [15], [22], [43]. However, the requirements of different applications such as news articles, social media, assistive technology, and so on, can be largely different. It remains difficult to tailor the generated image captions to a variety of such applications. The reason is manifold: First, image captioning approaches usually target to generate captions close to those in existing training data, and then are

The associate editor coordinating the review of this manuscript and approving it for publication was Hiu Yung Wong^{id}.

evaluated based on their similarity to the testing data. Both the datasets and the evaluation metrics are made under the assumption of performing general-purpose image captioning. This generally results in a very low diversity of generated captions, as some research has tried to tackle [9], [39], [41]. Second, the perception and the style of the generated captions are rarely considered, although some research looked into captioning styles and sentiment [3], [11], [24] and the visual descriptiveness of captions [36]. Recent research towards caption diversification propose introducing parameters such as length-controllable models [7].

In this paper, we explore the diverse generation of image captions with two controllable parameters: *imageability* and

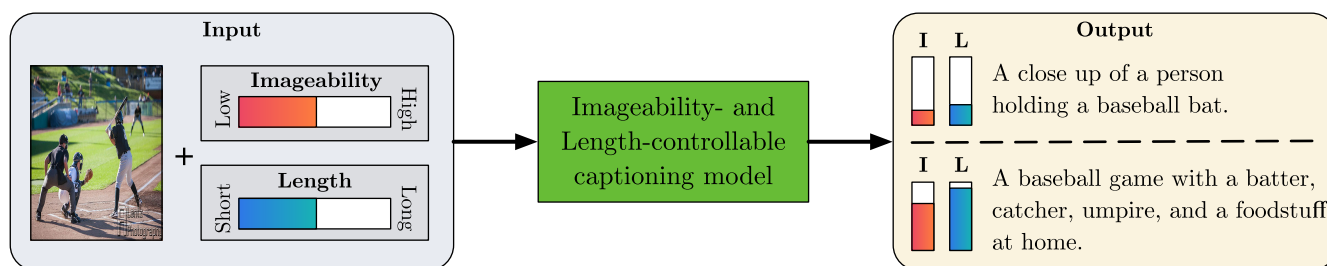


FIGURE 1. Proposed imageability- and length-controllable image captioning model. The *imageability* parameter allows for adjusting the visual descriptiveness on captions with the same length, while the *length* parameter changes the length for a fixed degree of visual descriptiveness. Both parameters can be changed at the same time to allow for creating diverse captions.

length. First, imageability, a concept derived from Psycholinguistics [27] which describes whether a word gives a clear mental image, is used. Its usage for image-captioning has been explored in our previous work [36], yielding promising results for customized image captions. In context of captioning, it can be used to adjust the *visual descriptiveness* of captions, making them being either a more abstract or more concrete description of the scene. Second, length provides another dimension of customizability of captions for different applications. While a news article might prefer a short abstract caption, a caption for assistive technology would be ideally longer and more descriptive. Further, by introducing two controllable variables, the proposed model can adjust both dimensions individually. The overall idea is illustrated in Fig. 1, showing how different settings for imageability and length can yield to vastly different captions. We believe that this step towards customized captioning can be a promising direction for application-tailored captioning.

This research is based on our previous work published in a conference proceedings [36]. This initial work showed promising results for imageability-aware captioning with an LSTM-based architecture, yet yielding a still mixed correlation to human perception and often unnatural captions. In this follow-up research, we employ a transformer-based captioning model [46] in order to greatly improve the naturalness of the results, making it more viable for actual use in targeting different applications. A data augmentation method similar to our previous work is used to diversify captions for visual descriptiveness. Furthermore, a length-controllable parameter [7] is newly introduced, in order to allow for adjusting the generated captions along a second dimension. With this, our combined model allows for changing customization across two dimensions independently. Note that imageability and length encode different things; Changing imageability aims to change visual descriptiveness of the caption for the same length, while length aims to change the wordiness while keeping contents similar. As such, we believe the proposed method, being able to control them individually, is a great first step towards tailoring captions to single applications with different needs of contents and descriptiveness. The evaluations show a greatly improved performance when generating customized captions, beating comparison methods. Especially, a crowd-sourced subjective evaluation shows a

significant improvement over our previous work [36], now closely correlating with the intended perception of the generated captions.

Our contributions can be summarized as follows:

- We propose an imageability- and length-controllable image captioning framework which can create diverse captions closely tailored to various applications.
- To the best of our knowledge, this is the first captioning framework which allows to adjust both imageability and length independently.
- The evaluation shows a significant improvement over our previous work for imageability-aware image captioning, partially due to the introduction of the transformer-based model.

II. RELATED WORK

In this section, we discuss related work regarding image captioning and imageability. The related work on image captioning can be categorized into general-purpose image captioning and affective image captioning. While the former simply tries to summarize an image in a short sentence, the latter puts focus on attributes like emotion/sentiment, style, user-feedback, or descriptiveness. A rough overview of the introduced work is visualized in Fig. 2.

A. GENERAL-PURPOSE IMAGE CAPTIONING

With the rise of deep learning-based models such as Long Short-Term Memory (LSTM) [14], general-purpose image captioning [16], [40], [43] achieved a great boost in performance.

More recently, transformer models [10], [37] using an attention mechanism have attracted researchers' attentions due to a very high performance in many natural language processing-related tasks. Following, many recent state-of-the-art models for image captioning [18], [46], [47] make use of a transformer-based architecture.

Zhou *et al.* [46] combine a transformer model with attention on visual features extracted from images [18], [32] for image captioning yielding very promising performance. Most recently, Cornia *et al.* [5] and Pan *et al.* [28] added more sophisticated attention modules to further improve the performance of transformer-based image captioning.

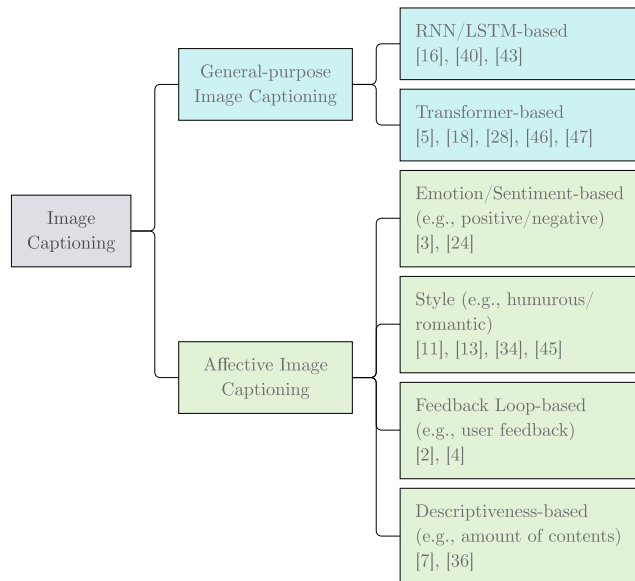


FIGURE 2. Related work in image captioning. The related work is split into general-purpose and affective image captioning. The former tries to simply summarize image contents in a neutral short phrase, while the latter puts a strong focus on the emotion/sentiment, style, feedback, descriptiveness, or other user perception of the output phrase.

B. AFFECTIVE IMAGE CAPTIONING

Rather than performing a neutral contents-based image captioning for general-purpose usage, there has been some research focus on image captioning in context of affective computing such as emotions and impressions [3]. They can be loosely categorized into four kinds of affective output:

First, Mathews *et al.* [24] propose a method which allows for customizing sentiment, yielding *positive* or *negative* sentiment captions.

Second, Gan *et al.* [11], Guo *et al.* [13], and Zhao *et al.* [45] explore the generation of styles such as *humorous* or *romantic*, which is further extended in a transformer-based model [34] to concepts like *sweet*, *dramatic*, *anxious*, *arrogant*, and so on.

Third, a different approach has been investigated by Cornia *et al.* [4], which allows user-interactive captioning where the user can specify image areas to be explained in a caption as well as their order. Chen *et al.* [2] propose similar ideas where scene graphs are used to fine-tune customized image captions.

Lastly, some approaches [7], [36] target specifying the detail and amount of output. Deng *et al.* [7] propose a length-controllable transformer model which can generate captions with fixed contents but a flexible length. In our previous work [36], we proposed a method for image captioning which can control the imageability of the generated captions. Imageability is a concept derived from Psycholinguistics first introduced by Paivio *et al.* [27], describing how easy it is to mentally imagine a word. It has received some attention in research for multi-modal analysis [25], [44], providing a promising opportunity to use it as a parameter for customized captioning.

In this research, we target the last discussed category of affective image captioning, proposing a method which allows for a high degree of customizability in descriptiveness of outputs. We build upon our previous work [36] on imageability-aware captioning using an LSTM-based model. We greatly improve the performance and naturalness of the generated captions by introducing a transformer-based captioning model [46]. As an additional parameter, we further introduce length-controllable captioning [7] to build a model which can generate captions with two independent parameters of customization.

III. IMAGEABILITY- AND LENGTH- CONTROLLABLE IMAGE CAPTIONING FRAMEWORK

In this section, we introduce the proposed framework for imageability- and length-controllable image captioning. For the imageability-controllable parameters, an augmented dataset with a high diversity in visual descriptiveness is needed. The augmentation and caption imageability estimation used in our method is largely based on our previous work [36], but briefly introduced in Sec. III-A due to this task being specialized and not yet receiving wide-spread attention. The proposed model itself is introduced in great detail in Sec. III-B.

A flowchart of the method is illustrated in Fig. 3.

A. DATASET PREPARATION

Following, we discuss the dataset needed for the proposed method. While the length-embedding of the framework is based on length-aware caption decoders as proposed by Deng *et al.* [7], the knowledge used for the imageability-embedding is trained on a diversified dataset. Thus, we first use a data augmentation technique to increase the number of captions in the dataset. The main focus lies on increasing the variety of visual descriptiveness of captions. Thus, we substitute information with more abstract terms, making captions more abstract for training. Next, the caption imageability is calculated for each caption, which is used for the imageability embedding during training.

1) DATA AUGMENTATION

Existing image captioning datasets such as Microsoft COCO [20] and Flickr30k [30] usually come with multiple captions for each image. However, there is typically not much diversity in terms of visual descriptiveness and each existing caption describes the image in a roughly similar way. For imageability-controllable captioning, we are interested in a large variety of descriptions, from abstract to visually descriptive. Imageability as a concept derived from Psycholinguistics [27] describes whether a word gives a clear mental image. For this research, we assume a rough relationship between visual descriptiveness and imageability, and thus use it to approximate a metric for visual descriptiveness. For a low target imageability, an ideal description would be something rather abstract, not mentioning many visual details.

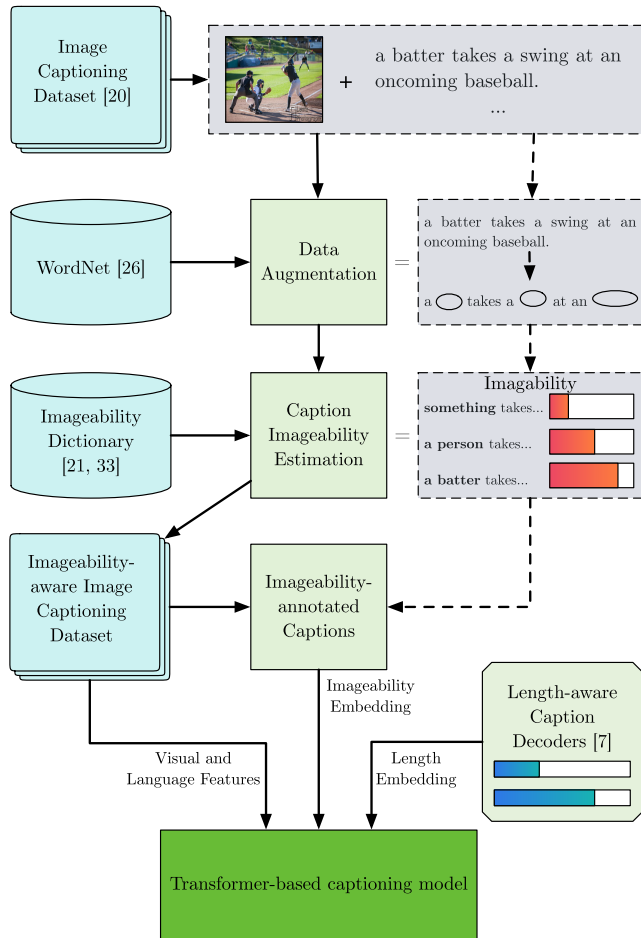


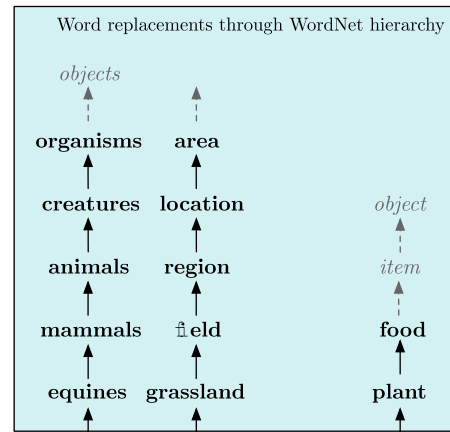
FIGURE 3. Flowchart of the proposed framework. A general purpose image captioning dataset is augmented using word substitutions through WordNet. This generates a diverse caption-dataset with different levels of visual descriptiveness. For each caption, an imageability score is calculated, which is then used for generating an imageability-embedding. The proposed model incorporates both an imageability- and a length-based embedding. The model itself is shown in Fig. 5.

In contrast, for a high target imageability, a very detailed description of visual details in the caption would be expected.

To emulate this idea, the augmentation process substitutes words in existing captions with more abstract terms. With the help of the transformer architecture, the augmented data can then help the network to identify abstract language and how it would change captions. Similar to our previous work [36], each noun in a given caption is substituted by their hypernym according to its WordNet [26] hierarchy. We replace a noun with up to five levels of hypernyms in order to generate additional captions. Note, that we avoid going too close to the WordNet root node by removing the top-most two layers, as terms like *object* or *item* become too abstract for meaningful training. For captions with multiple nouns, we generate augmented captions for each noun separately. The idea is visualized in Fig. 4.

2) CAPTION IMAGEABILITY ESTIMATION

In order to learn the relationship between an image and the visual descriptiveness of a caption, we calculate the caption



Two brown horses in a pasture are eating the grass.

FIGURE 4. Data augmentation. Using WordNet [26], we extract a hierarchy of hypernym terms for each noun in the existing captions. We pick up to five replacements for each noun, e.g., replacing *pasture* with the terms {*area, location, region, field, grassland*}. Note that we avoid replacements too close to the WordNet root node, as they would become too abstract. As such, *grass* will only be augmented by {*food, plant*}, but not with *item* or *object* which would come above. This process is repeated for all nouns in every caption to create an augmented dataset with more abstract wordings.

imageability. The basic idea is to use imageability values for individual words composing the caption in order to calculate a value representative for the whole caption. Existing imageability dictionaries such as [6], [31], [33], [42] describe imageability on a Lickert scale (e.g., on an interval of [1,7] or [1,5]) from very unimaginable to very imaginable.

For caption-imageability estimation, we follow the same approach as in our previous work [36]. We start with a caption from the dataset and assume available imageability labels for all its individual words. As this is a strong assumption, we skip stop-words, numerals, and the similar. For our experiments, we target English language, which also influences some design decisions discussed onwards, but an adjusted process is expected to work for other languages, too. We generate a parsing tree using the Stanford CoreNLP [23] framework. Next, we employ a bottom-up approach which calculates a sentence imageability score from all its words' imageability values along the parsing tree. We assume nouns to become more descriptive when being modified by adjectives (e.g., "black cat" being a less visually ambiguous description than "cat"). For multiple words on the same level of the parsing tree, we define some simple rule set for weighting: 1) If there are one or more nouns, the last noun is the most significant and weighted the highest (e.g., "cold apple juice" are modifications of "juice"). 2) If there is no noun, the first word is the most significant and weighted the highest (e.g., "run fast" is a modification of "run"). We calculate the imageability of sub-trees using

$$I = x_s \prod_{i=1(\neq s)}^n (2 - e^{-x_i}), \tag{1}$$

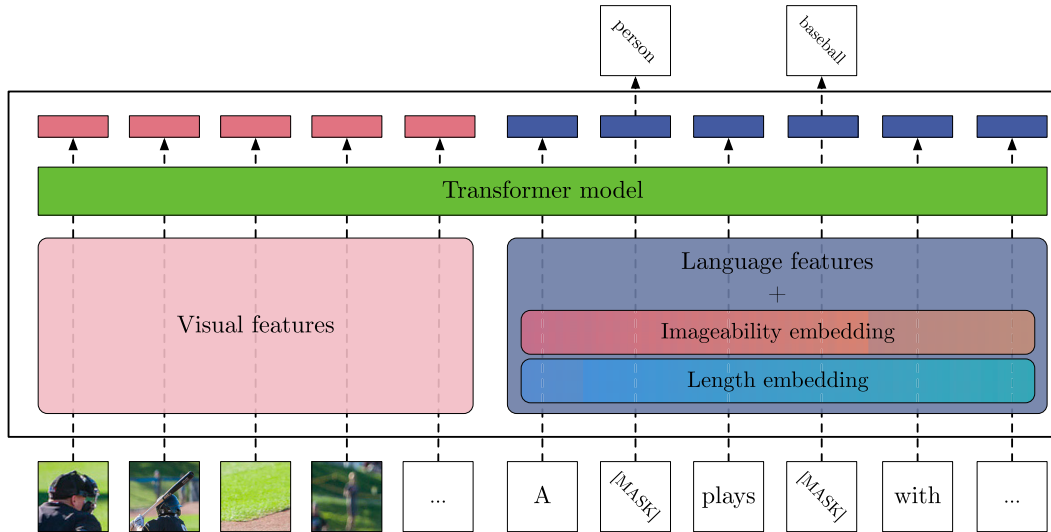


FIGURE 5. Proposed captioning model. The proposed model uses a transformer-based architecture. It is based on [7] which allows for length-controllable captioning. Inspired by their architecture, the proposed methods adds an imageability embedding layer which encodes the visual descriptiveness of captions. Using this, the resulting model allows both imageability- and length-controllable output.

where x_i ($i = 1, \dots, n \mid i \neq s$) is the score of each modifying word and x_s is the score of the most significant word. This process is repeated bottom-up until reaching the root node of the parsing tree. Lastly, the results are normalized using $f(x) = 1 - e^{-x}$.

We employ this method and calculate the caption imageability values for all captions in the augmented dataset.

B. CAPTIONING MODEL

For the captioning model, we employ a BERT-based transformer model [46]. Deng *et al.* [7] apply this model for length-controllable captioning, where they add a layer of length-embedding to the language features. Inspired by this, we add an extra layer of imageability-embedding based on the augmented dataset with caption imageability estimations. Our proposed model is illustrated in Fig. 5.

First, we introduce each type of embedding and the features used for the training.

1) LENGTH EMBEDDING

The length embedding is implemented in the same fashion as proposed by Deng *et al.* [7].

For a caption $C = \{c_i\}_{i=1}^N$, with c_i representing each word in a caption, we assign C a length level with the range $[L_{low}, L_{high}]$ according to its length N . Then, the length-embedding matrix $W_l \in \mathbb{R}^{k \times d}$ (with k being the number of length levels and d being the embedding dimension) is trained to differentiate image captions on different length levels.

A one-hot vector $\mathbf{t}_l \in \mathbb{R}^d$ for the length l is generated. The length embedding is then defined as

$$\mathbf{e}_{len} = W_l^T \mathbf{t}_l \in \mathbb{R}^d. \quad (2)$$

2) IMAGEABILITY EMBEDDING

Inspired by the length embedding discussed before, we implement an imageability embedding in the same way. For each caption, we generate an imageability embedding based on the caption imageability estimation obtained in Sec. III-A. We assign an imageability level i to a caption within a range of $[I_{low}, I_{high}]$ according to its caption imageability I . Through this, the existing caption imageability annotations are binned into evenly-sized levels. The imageability-embedding matrix $W_i \in \mathbb{R}^{a \times d}$ (with a being the number of imageability levels and d being the embedding dimension) is trained to differentiate image captions on different imageability levels. $\mathbf{t}_i \in \mathbb{R}^a$ represents a one-hot vector for the imageability level. Finally, the imageability embedding becomes

$$\mathbf{e}_{imag} = W_i^T \mathbf{t}_i \in \mathbb{R}^d. \quad (3)$$

3) VISUAL FEATURES

The model applies a Faster-RCNN [32] network pre-trained on the Visual Genome dataset [17] to extract visual features. Using this object detection model, the regions $R = \{r_i\}_{i=1}^M$ corresponding to M objects are detected. We extract region features $\mathbf{F}_e = \{\mathbf{f}_{e,i}\}_{i=1}^M$, classification probabilities $\mathbf{F}_c = \{\mathbf{f}_{c,i}\}_{i=1}^M$, and localization features $\mathbf{F}_l = \{\mathbf{f}_{l,i}\}_{i=1}^M$ for each object in the image.

The visual features are then defined as

$$\mathbf{x}_{r_i} = W_e^T \mathbf{f}_{e,i} + W_p^T [LN(\mathbf{f}_{c,i}), LN(\mathbf{f}_{l,i})] + \mathbf{e}_{vis}, \quad (4)$$

describing the visual vector \mathbf{x}_{r_i} for the region r_i . Here, \mathbf{e}_{vis} is a learnable embedding for differentiating the image regions from text tokens. The projection matrices W_e and W_p are trainable and project the corresponding features into d -D space. LN refers to layer normalization while $[\cdot, \cdot]$ represents feature vector concatenation.

4) LANGUAGE FEATURES

For an input caption $C = \{c_i\}_{i=1}^N$, we use a BERT-based model [46] to obtain a word-embedding $\mathbf{e}_{w,c_i} \in \mathbb{R}^d$ and a location-embedding $\mathbf{e}_{p,i} \in \mathbb{R}^d$.

The length- and imageability-embeddings are added to the language features, which are defined as

$$\mathbf{x}_{c_i} = \mathbf{e}_{w,c_i} + \mathbf{e}_{p,i} + \mathbf{e}_{len} + \mathbf{e}_{imag}. \quad (5)$$

5) MODEL TRAINING

The proposed model is based on the language generation model by Ghazvininejad *et al.* [12]. For a correct caption $T = \{t_i\}_{i=1}^N$, which is randomly masked with tokens $[MASK]$, the transformer network is fed with a masked caption $C = \{c_i\}_{i=1}^N$. Next, the pair of visual and language features is fed into the network, predicting the masked token. The model is trained by minimizing the cross-entropy loss between the correct token t_i of the ground-truth caption and the masked-in token c_i as expressed by

$$L = - \sum_{i=1}^N \mathbb{1}(c_i) t_i \log c_i. \quad (6)$$

Note that $c_i = [MASK]$ is an indicator function that is 1 only when $\mathbb{1}(\cdot)$, and 0 otherwise.

6) CAPTION GENERATION

Following Ghazvininejad *et al.* [12], we use the ‘‘Mask-Predict-Update’’ method to generate captions. Initially, the whole caption is masked with $[MASK]$ tokens. The feature embeddings are fed into the transformer network in order to predict a mask position and its most suitable vocabulary. The process is repeated iteratively until the whole caption is generated.

IV. EVALUATION

In this section, we evaluate our proposed image captioning method. After discussing the environment in Sec. IV-A, we illustrate some generated captions of the proposed method in Sec. IV-B.

Following, we evaluate the approach from three angles: First, Sec. IV-C discusses the performance of the model measured by general-purpose image captioning metrics. The length-controllable transformer-based method has already been extensively evaluated in [7]. Therefore, for the second and third experiments, we focus on a deeper evaluation of the imageability-controllable part of the transformer-based model and its differences over the previous LSTM-based work [36] for generating captions with different visual descriptiveness. As such, Sec. IV-D discusses the imageability diversity of the generated captions, and Sec. IV-E the performance in a crowd-sourced human evaluation.

A. ENVIRONMENT

1) DATASETS

We employ the Microsoft COCO [20] dataset as a baseline for the data augmentation. For training and testing, we use

Karpathy splits [16]. The extended dataset is generated as discussed in Sec. III-A1, aiming for twenty captions per image. For the imageability estimation of captions, we employ two imageability dictionaries by Ljubešić *et al.* [21] and Scott *et al.* [33]. As the former is a large estimated dictionary while the latter is a small crowd-sourced one, we favor the ground-truth imageability of the latter dictionary in case of overlaps. Images which did not yield sufficient numbers of captions through data augmentation or did not have sufficient imageability word annotations were excluded from the experiments. We end up with 109,115 images for training, 4,819 images for validation, and 4,795 images for testing.

2) IMPLEMENTATIONS

We use a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [10] model consisting of twelve layers of transformers. For both imageability and length, we define classes as discussed in Sec. III.

For the imageability-controllable parameter, we define five levels of imageability. The imageability from dictionaries is normalized to an interval of $[0, 1]$. Due to the distribution of imageability values in the original datasets, virtually all captions result in an imageability above 0.5 through the method discussed in Sec. III-A2. Thus, splitting the resulting data evenly, we end up with the five imageability levels: I-1 (imageability between $(0.5, 0.6]$), I-2 ($(0.6, 0.7]$), I-3 ($(0.7, 0.8]$), I-4 ($(0.8, 0.9]$), and I-5 ($(0.9, 1.0]$) used for training. For the experiment, we are interested in how the imageability captures human perception, i.e., whether the visual descriptiveness of different levels actually resemble the expectations of a human. As neighboring imageability levels are very close and sometimes perceptually overlap, we evaluate three classes in order to understand the overall trend of results —concretely choosing: Low (I-1), Mid (I-3), and High (I-5).

For the controllable length parameter, we define four length levels: L-1 (length of $[7, 9]$ with 10 iterations of Mask-Predict-Update), L-2 ($[10, 14]$, 15 iterations), L-3 ($[15, 19]$, 20 iterations), and L-4 ($[20, 24]$, 25 iterations).

We evaluate all combinations of L- x and I- x regarding their qualitative and quantitative results. We furthermore also evaluate a variant where we only use the imageability-controllable features I- x and exclude the length-embedding. The reason for this is that the length-controllable transformer model have been already exhaustively evaluated in [7], while the imageability-controllable part of the transformer model is a contribution of this paper.

3) COMPARISON METHODS

For comparison, we tested a selection of methods from related work on the same datasets.

First, we want to understand how the performance of our imageability- and length-controllable captioning method compares to general-purpose captioning. Thus, in Sec. IV-C, we compare our results to a general-purpose method, ‘‘Show, Attend, and Tell’’ (SAT) by Xu *et al.* [43], the

TABLE 1. Example of generated image captions when changing the target imageability and the length at the same time. The results verify a promising performance for generating diverse captions for different applications.





Image	Length level	Imageability level	Caption
	L-1	Low	Some organisms are playing in a baseball game.
		Mid	A batter taking a mechanism at a ball.
		High	A baseball person at bat during a game.
	L-2	Low	A close up of a person holding a baseball bat.
		Mid	A male is swinging a bat at a baseball game.
		High	A baseball person holding a bat on a field.
	L-3	Low	A close up of a baseball player holding a vertebrate on a field.
		Mid	A foodstuff getting ready to swing at a ball during the game.
		High	A baseball person holding a bat with a catcher and umpire standing behind him.
	L-4	Low	A close up of a baseball player holding a vertebrate with a catcher and umpire behind him.
		Mid	A foodstuff getting ready to hit, while the catcher is getting ready to catch the ball.
		High	A baseball game with a batter, catcher, umpire, and a foodstuff at home plate.

TABLE 2. Example captions as qualitative comparison. As TAYI [36] cannot generate length-aware captions, these examples use the proposed method without the length embedding. The results show that the proposed method generates much more natural results for the same imageability setting, and a higher variety of descriptiveness in general (bold highlights).

Image	Method	Imageability level	Caption
	TAYI [36]	Low	A placental is sitting on a window sill.
		Mid	A feline is sitting on a window sill.
		High	A cat is sitting on a window sill.
	Proposed	Low	A close up of a cat near a glass window sill.
		Mid	A vertebrate is looking out of a window.
		High	A brown and white cat sitting on a window sill.
	TAYI [36]	Low	A large brown canine laying on top of a beach.
		Mid	A large brown canine laying on top of a beach.
		High	A large brown dog laying on top of a beach.
	Proposed	Low	A close up of a canine laying on a beach.
		Mid	A carnivore laying on the ground in the sand .
		High	A brown and white dog laying on a beach.
	TAYI [36]	Low	An organism swinging a baseball bat at a baseball .
		Mid	An organism swinging a baseball bat during a baseball game .
		High	A baseball player swinging a bat at a ball .
	Proposed	Low	A concoction getting ready to swing at a pitch .
		Mid	A male is up to bat during a baseball game .
		High	A baseball person holding a bat on a field .

length-controllable approach LaBERT by Deng *et al.* [7] (using their best-performing variant with L-2 for the comparison), as well as general-purpose methods X-Transformer by Pan *et al.* [28] and M^2 by Cornia *et al.* [5].

Second, we include our previous work “Tell As You Imagine” (TAYI) [36], which generates imageability-aware captions using an LSTM-based approach. This work is not trained on grouped imageability levels, but can generate individual values of imageability $I = [0.5, 0.6, \dots, 0.9]$. To yield a comparable output, similar to the way we defined levels in the proposed method, we generate captions for Low (with $I = 0.5$), Mid ($I = 0.7$), and High ($I = 0.9$). We use this as the main comparison method for experiments in Sec. IV-D and IV-E, as it is to the best of our knowledge, the only related work tailoring its output to imageability.

B. QUALITATIVE EVALUATION

Before looking into the quantitative metrics, we showcase some examples of the output of the proposed method. Table 1 shows the output for an example image where imageability- and length-parameters were adjusted at the

same time. We can see that the customization works well in both dimensions, allowing for a promising way to tailor the model output to individual needs of applications. Note that this also results in a high caption diversity which could also be useful for many applications. To the best of our knowledge, there is no other method which can generate both imageability- and length-controllable captions. Thus, we cannot provide a comparison method.

TAYI [36] is the only related work targeting imageability-aware captioning. We compare it to our proposed model in Table 2. In this case, we excluded the length-embedding, resulting in results which roughly resemble those of length level L-2. As we can see here, the output of our method vastly outperforms this comparison method, making the results much more natural. This is mostly a result from the switch to a transformer-based architecture compared to LSTM used in the comparison method.

For length-controllable captions, LaBERT [7] provides an exhaustive analysis. As our architecture without the imageability embedding is largely identical to their setup, we thus skip a more detailed analysis of this parameter.

TABLE 3. Evaluation through general-purpose image captioning metrics. The proposed method is compared to TAYI [36] which is the only other related work aiming at imageability-aware captioning and [5], [7], [28], [43] in order to compare performance against general-purpose captioning models. Due to the very different style of captions generated for different levels of imageability, the scores are split into three groups, highlighting the average performance for a low, mid, and high target imageability. The bold values correspond to the highest value within the imageability-aware methods.

Method	BLEU-4 [29]			CIDEr [38]			ROUGE [19]			METEOR [8]			SPICE [1]		
	Low-I	Mid-I	High-I	Low-I	Mid-I	High-I	Low-I	Mid-I	High-I	Low-I	Mid-I	High-I	Low-I	Mid-I	High-I
Imageability-aware image captioning															
TAYI [36]	0.265	0.262	0.246	0.621	0.633	0.618	0.495	0.495	0.491	0.232	0.235	0.238	0.089	0.092	0.093
Prop. (I)	0.247	0.294	0.290	0.671	0.747	0.850	0.488	0.536	0.538	0.234	0.255	0.264	0.094	0.101	0.110
Prop. (I+L-1)	0.222	0.263	0.241	0.553	0.654	0.714	0.459	0.518	0.511	0.208	0.232	0.236	0.080	0.088	0.096
Prop. (I+L-2)	0.248	0.295	0.289	0.683	0.758	0.850	0.489	0.537	0.540	0.234	0.255	0.264	0.094	0.101	0.108
Prop. (I+L-3)	0.205	0.231	0.240	0.589	0.633	0.712	0.472	0.503	0.513	0.244	0.260	0.272	0.102	0.107	0.116
Prop. (I+L-4)	0.166	0.181	0.184	0.316	0.342	0.360	0.433	0.451	0.460	0.246	0.257	0.265	0.109	0.112	0.121
General purpose image captioning															
SAT [43]		0.281			0.671			0.504			0.238			0.092	
LaBERT [7]		0.328			0.895			0.560			0.273			0.110	
X-Trans. [28]		0.372			1.204			0.576			0.287			0.218	
M ² [5]		0.393			1.318			0.587			0.293			0.226	

C. EVALUATION WITH IMAGE CAPTIONING METRICS

For this experiment, we evaluate our proposed method against comparison methods [5], [7], [28], [36], [43] regarding the general-purpose image captioning metrics BLEU [29], CIDEr [38], ROUGE [19], METEOR [8], and SPICE [1]. The results are shown in Table 3. As general-purpose image captioning and imageability-aware image captioning are strictly speaking different tasks and not directly comparable, we grouped these methods for better visibility.

Overall, the imageability-aware models yield a reasonable performance across all metrics, despite the more recent general-purpose methods outperforming them. As the proposed method discusses a specialized task of imageability- and length-controllable captioning, we did not expect to achieve the best performance in these metrics. Rather than performing the best, we want to aim for a reasonable performance while providing an additional dimension of customizability. Note that most of the evaluation metrics actually do not consider, but rather punish, diverse captions and style changes, as the evaluation is based on a direct comparison to a ground-truth annotation. As such, methods aiming for diversification or affective computing commonly slightly degrade performance in such metrics by their nature. The method by LaBERT [7] outperformed our proposed method in most metrics, but the results are close enough to verify a similar performance. As we were interested in general-purpose performance, we used the best-performing variant (L-2) of their model.

Newer architectures such as [5], [28] further outperform the proposed method. Because of this, future research could investigate into whether these architectures could also be beneficial for imageability-aware captioning.

Note that the nature of the approach, actively purposefully changing contents of the output, would naturally *decrease* their performance in terms of these general-purpose image captioning metrics.

We can also see a great improvement over TAYI [36], which also aimed for imageability-aware captioning. Here, the proposed method outperformed the comparison method on all metrics.

TABLE 4. Quantitative evaluation of imageability-controllable captions. The proposed method is compared to TAYI [36] which is the only other related work aiming at imageability-aware captioning. This table shows the output range of the proposed model. The variety and imageability range are indicators for the diversity of the generated captions. Note that the root mean squared error (RMSE) is not directly comparable as the comparison method is trained on discrete imageability values on an interval of [0, 1] while the proposed method is trained on five imageability levels (changing the interval to [0, 4]).

Method	Caption variety	Imag. range	RMSE		
			Low-I	Mid-I	High-I
TAYI [36]	2.755	0.091	0.274	0.107	0.084
Proposed (I)	4.827	0.335	A0.438	0.329	0.181
Proposed (I+L-1)	4.723	0.343	0.290	0.258	0.142
Proposed (I+L-2)	4.849	0.335	0.441	0.348	0.183
Proposed (I+L-3)	4.848	0.334	0.704	0.543	0.196
Proposed (I+L-4)	4.924	0.326	1.162	0.726	0.179

D. EVALUATION OF IMAGEABILITY-CONTROLLABLE CAPTIONS

In this experiment, we evaluate the imageability-controllable captions. Here, we analyze the variety of the generated captions.

The results are shown in Table 4. We can see that the proposed method is able to yield an overall increased variety of captions. While TAYI [36] aims for generating individual results for imageability between [0.5, 0.6, . . . , 0.9], most will actually result in very similar or identical captions. Similarly, the range of output imageability is rather compact. In contrast, the proposed method can generate a higher variety of diverse captions, yielding up to five distinct captions (i.e., usually having individual results for each imageability level I-1 to I-5). Furthermore, the span of imageability is higher, leading to a perceptually larger difference between the generated captions.

E. SUBJECTIVE EVALUATION

Lastly, in this section, we explore the human perception of the generated captions. As the imageability-controlled captions are expected to have a varying degree of visual descriptiveness, we are interested in whether this intended effect matches the perception of users when reading the caption. Following, we performed a crowd-sourced subjective evaluation where

we asked participants to judge pairs of captions regarding how easy they are to visually imagine. Note that we do not include other related methods such as SAT [43] in the comparison, as those methods provide no meaningful way to generate multiple captions with different perceptions (such as visual descriptiveness). As such, we compare our results only to TAYI [36], which is the only related work with such a parameter.

We generated three English captions each for 195 images, corresponding to the Low (I-1), Mid (I-2), and High (I-5) imageability levels as discussed before. Using Amazon Mechanical Turk¹ we asked participants to perform a Thurstone's paired comparison task [35], judging which caption is easier to visually imagine based on its textual contents. Note that we do not show the actual image, because we also want to see whether a high imageability might help making a caption more suitable for assistive technologies. For each pair, we asked fifteen US participants to obtain a meaningful majority decision. The human judgements were compared to the intended imageability values using Pearson's rank correlation. The results are shown in Table 5. The values in the right-half of the table show the distribution of fully matching, half-matching, inverse-half-matching and inverse-fully-matching between our intended imageability and human perception. The avg. column shows the overall correlation for each method. The proposed method vastly outperformed the comparison method, resulting in an average correlation of 0.70 over a correlation of 0.36 in the comparison method. Note that the 95% CI column shows 95% confidence intervals for each method. As discussed before, TAYI uses an LSTM-based architecture while the proposed method uses a transformer-based architecture, resulting in a well-improved performance. Together with the more natural results illustrated in Table 1, we believe that the proposed method provides a meaningful framework useful for many real-world applications.

TABLE 5. Subjective evaluation of visual descriptiveness. The proposed method is compared to TAYI [36] which is the only other related work aiming at imageability-aware captioning. In the survey, participants were asked to judge the mental image of a pair of captions. The results show the correlation between the human perception of generated captions and the target-imageability. For this experiment, the length embedding is excluded, using only the imageability-controllable setting.

Method	ρ					
	Avg.	95% CI	-1.0	-0.5	0.5	1.0
TAYI [36]	0.36	[-0.19, 0.74]	0.05	0.22	0.43	0.30
Proposed	0.70	[0.29, 0.89]	0.01	0.03	0.46	0.50

V. CONCLUSION

In this paper, we proposed a transformer-based method to generate diverse image captions with two controllable dimensions: First, building upon our previous work on imageability-aware captioning, TAYI [36], we use *imageability* as a parameter to change the degree of visual

descriptiveness of a generated caption. Second, inspired by recent work on length-controllable captioning [7], we use *length* as another parameter to modify the length of a caption independent of the degree of visual descriptiveness. Imageability and length encode two different angles: Changing imageability aims to change visual descriptiveness of the caption for the same length, while length aims to change the wordiness while keeping contents similar. This capability allows to tailor the output captions towards different use-cases for accessibility reasons, different media, or different user preferences. The resulting model is, to the best of our knowledge, the first model which can generate a variety of differently-perceived captions tailored to various applications.

In the experiments, the proposed method showed a promising performance for generating captions across different lengths and imageability values. A subjective evaluation with human participants verified a vastly improved performance compared to an existing method. This shows that the transformer architecture in combination with imageability as a prior can successfully learn the human perception of sentences regarding the degree of visual descriptiveness. For future work, it could be interesting to look into other Transformer-based architectures such as [5], [28].

REFERENCES

- [1] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Computer Vision (Lecture Notes in Computer Science)*, vol. 9909, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 382–398.
- [2] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 9962–9971.
- [3] T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin, and J. Luo, "'Factual' or 'emotional': Stylized image captioning with adaptive learning and attention," in *Computer Vision (Lecture Notes in Computer Science)*, vol. 11214, M. Hebert, C. Sminchisescu, Y. Weiss, Eds., Munich, Germany. Cham, Switzerland: Springer, Sep. 2018, pp. 527–543.
- [4] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8307–8316.
- [5] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.
- [6] M. J. Cortese and A. Fugett, "Imageability ratings for 3,000 monosyllabic words," *Behav. Res. Methods, Instrum., Comput.*, vol. 36, no. 3, pp. 384–387, Aug. 2004.
- [7] C. Deng, N. Ding, M. Tan, and Q. Wu, "Length-controllable image captioning," in *Computer Vision (Lecture Notes in Computer Science)*, vol. 12358, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 712–729.
- [8] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, Baltimore, MD, USA, Jun. 2014, pp. 376–380.
- [9] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 10695–10704.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Minneapolis, MN, USA, vol. 1, Jun. 2019, pp. 4171–4186.

¹<https://www.mturk.com/>

- [11] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3137–3146.
- [12] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Hong Kong, 2019, pp. 6112–6121.
- [13] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "MSCap: Multi-style image captioning with unpaired stylized text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 4204–4213.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 4565–4574.
- [16] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3128–3137.
- [17] 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 L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
- [18] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 8928–8937.
- [19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop ACL Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81.
- [20] 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 *Computer Vision (Lecture Notes in Computer Science)*, vol. 8693, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Zurich, Switzerland: Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.
- [21] N. Ljubešić, D. Fišer, and A. Peti-Štantić, "Predicting concreteness and imageability of words within and across languages via word embeddings," in *Proc. 3rd Workshop Represent. Learn.*, Melbourne, VIC, Australia, Jul. 2018, pp. 217–222.
- [22] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 7219–7228.
- [23] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Ann. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Baltimore, MD, USA, Jun. 2014, pp. 55–60.
- [24] A. P. Mathews, L. Xie, and X. He, "SentiCap: Generating image descriptions with sentiments," in *Proc. 30th AAAI Conf. Artif. Intell.*, Palo Alto, CA, USA, vol. 30, Feb. 2016, pp. 3574–3580.
- [25] C. Matsuhira, M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, K. Doman, D. Deguchi, and H. Murase, "Imageability estimation using visual and language features," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 306–310.
- [26] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [27] A. Paivio, J. C. Yuille, and S. A. Madigan, "Concreteness, imagery, and meaningfulness values for 925 nouns," *J. Experim. Psychol.*, vol. 76, no. 1, pp. 1–25, 1968.
- [28] Y. W. Pan, T. Yao, Y. H. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10971–10980.
- [29] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318.
- [30] 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. ICCV*, Santiago, Chile, Dec. 2015, pp. 2641–2649.
- [31] J. Reilly and J. Kean, "Formal distinctiveness of high and low-imageability nouns: Analyses and theoretical implications," *Cognit. Sci.*, vol. 31, no. 1, pp. 157–168, Feb. 2007.
- [32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Neural Information Processing Systems*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., Montreal, QC, Canada: Red Hook, NY, USA: Curran Associates, Dec. 2015.
- [33] G. G. Scott, A. Keitel, M. Becirspahic, B. Yao, and S. C. Sereno, "The Glasgow norms: Ratings of 5,500 words on nine scales," *Behav. Res. Methods*, vol. 51, no. 3, pp. 1258–1270, Jun. 2019.
- [34] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 12516–12526.
- [35] L. L. Thurstone, "The method of paired comparisons for social values," *J. Abnormal Social Psychol.*, vol. 21, no. 4, pp. 384–400, 1927.
- [36] K. Umemura, M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase, "Tell as you imagine: Sentence imageability-aware image captioning," in *MultiMedia Modeling (Lecture Notes in Computer Science)*, vol. 12573, J. Lokoč, T. Skopal, K. Schoeffmann, V. Mezaris, X. Li, S. Vrochidis, and I. Patras, Eds., Prague, Czech Republic: Cham, Switzerland: Springer, Jan. 2021, pp. 62–73.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds., 2017, pp. 5998–6008.
- [38] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 4566–4575.
- [39] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search for improved description of complex scenes," in *Proc. 32nd AAAI Conf. Artif. Intell.*, vol. 32, Palo Alto, CA, USA, Feb. 2018, pp. 7371–7379.
- [40] 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.*, Boston, MA, USA, Jun. 2015, pp. 3156–3164.
- [41] Q. Wang and A. B. Chan, "Describing like humans: On diversity in image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4190–4198.
- [42] M. Wilson, "MRC psycholinguistic database: Machine-usable dictionary, version 2.00," *Behav. Res. Methods, Instrum., Comput.*, vol. 20, no. 1, pp. 6–10, Jan. 1988.
- [43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Lille, France, Jul. 2015, pp. 2048–2057.
- [44] M. Zhang, R. Hwa, and A. Kovashka, "Equal but not the same: Understanding the implicit relationship between persuasive images and text," in *Proc. Brit. Mach. Vis. Conf.*, England, U.K., Sep. 2018, pp. 1–14.
- [45] W. Zhao, X. Wu, and X. Zhang, "MemCap: Memorizing style knowledge for image captioning," in *Proc. 34th AAAI Conf. Artif. Intell.*, vol. 34, Palo Alto, CA, USA, Apr. 2020, pp. 12984–12992.
- [46] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, Palo Alto, CA, USA, Apr. 2020, vol. 34, no. 7, pp. 13041–13049.
- [47] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning transformer with stacked attention modules," *Appl. Sci.*, vol. 8, no. 5, pp. 739–749, May 2018.



MARC A. KASTNER received the B.Sc. and M.Sc. degrees in computer science from the Braunschweig University of Technology, Brunswick, Germany, in 2013 and 2016, respectively, and the Ph.D. degree in informatics from the Graduate School of Informatics, Nagoya University, Japan, in 2020.

Since 2020, he has been working as a Post-doctoral Researcher at the National Institute of Informatics, Japan. His research interests include the connection of the human with multimedia, covering vision & language- and affective computing-related tasks.

Dr. Kastner is a member of ACM and IPS Japan.



KAZUKI UMEMURA received the B.Eng. and M.S. degrees from Nagoya University, Japan, in 2019 and 2021, respectively.

After finishing this research, he joined NEC Corporation, Japan. His research interests include computer vision and natural language processing, focusing on the use of psycholinguistic features in image captioning.



ICHIRO IDE (Senior Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees from The University of Tokyo, in 1994, 1996, and 2000, respectively.

He became an Assistant Professor at the National Institute of Informatics, Japan, in 2000, and an Associate Professor at Nagoya University, Japan, in 2004, where he has been a Professor, since 2020. He was a Visiting Associate Professor at the National Institute of Informatics,

from 2004 to 2010, an Invited Professor at the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), France, in 2005, 2006, and 2007, and a Senior Visiting Researcher at ISLA, Instituut voor Informatica, Universiteit van Amsterdam, from 2010 to 2011. His research interests include the analysis and indexing to authoring and generation of multimedia contents, especially in large-scale broadcast video archives and social media, mostly on news, cooking, and sports contents.

Dr. Ide is a Senior Member of IEICE and IPS Japan, and a member of ACM, JSAI, and ITE.



YASUTOMO KAWANISHI (Member, IEEE) received the B.Eng. degree in engineering and the M.Inf. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. He became a Postdoctoral Fellow at Kyoto University, in 2012. In 2014, he moved to Nagoya University, Japan, as a Designated Assistant Professor, where he became an Assistant Professor, in 2015, and a Lecturer, in 2020. Since 2021, he has been the Team Leader of the

Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His main research interests include robot vision for environmental understanding and computer vision for human understanding, especially pedestrian detection, tracking, retrieval, and recognition. He is a member of IEEEJ and IEICE. He received the Best Paper Award from SPC2009 and the Young Researcher Award from the IEEE ITS Society Nagoya Chapter.



TAKATSUGU HIRAYAMA (Member, IEEE) received the M.E. and D.E. degrees in engineering science from Osaka University, in 2002 and 2005, respectively.

From 2005 to 2011, he was a Research Assistant Professor with the Graduate School of Informatics, Kyoto University, Japan. In 2011, he moved to the Graduate School of Information Science, Nagoya University, Japan, where he became an Assistant Professor, in 2012,

and a Designated Associate Professor, in 2014. In 2017, he became a Designated Associate Professor at the Institutes of Innovation for Future Society, Nagoya University. Since 2021, he has been a Professor with the University of Human Environments, Japan. His research interests include computer vision (face recognition, visual attention modeling, and action recognition) and human-computer interaction (multi-modal interaction design, internal state estimation, and interaction dynamics analysis).

Dr. Hirayama is a member of IEICE, IPS Japan, and ACM.



KEISUKE DOMAN (Member, IEEE) received the B.S. degree from the Department of Electrical and Electronic Engineering, Nagoya University, Japan, in 2007, and the M.S. and Ph.D. degrees from the Graduate School of Information Science, Nagoya University, in 2009 and 2012, respectively.

In 2013, he became a Lecturer at the School of Information Science and Technology, Chukyo University, Japan, where he has been an Associate Professor with the School of Engineering, since

2020. His current research interests include the application of computer vision and pattern recognition to human activity support systems.

Dr. Doman is a member of IEICE, ACM, and INSTICC.



DAISUKE DEGUCHI (Member, IEEE) received the B.Eng. and M.Eng. degrees in engineering and the Ph.D. degree in information science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively.

He is currently an Associate Professor with the Graduate School of Informatics, Nagoya University. He is working on object detection, segmentation, and recognition from videos, and their applications to ITS technologies, such as detection

and recognition of traffic signs.



HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist at the Columbia University in the City of New York, New York. He has been a Professor with Nagoya University, since 2003. His research interests include computer vision, pattern

recognition, and multimedia information processing. He is a fellow of the IPSJ and the IEICE. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He received the Medal with Purple Ribbon from the Government of Japan, in 2012.



SHIN'ICHI SATOH (Member, IEEE) received the B.E. degree in electronics engineering and the M.E. and Ph.D. degrees in information engineering from The University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively.

He was a Visiting Scientist at the Robotics Institute, Carnegie Mellon University, Pittsburgh, USA, from 1995 to 1997. He has been a Full Professor with the National Institute of Informatics, Tokyo, since 2004. His current research interests

include image processing, video content analysis, and multimedia databases.

Dr. Satoh is a member of ACM, IEICE, and IPS Japan.

...