

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

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^{5,2}, (MEMBER, IEEE), KEISUKE DOMAN⁶, (MEMBER, IEEE), DAISUKE DEGUCHI², (MEMBER, IEEE), HIROSHI MURASE², (FELLOW, IEEE), AND SHIN'ICHI SATOH¹, (MEMBER, IEEE).

¹Digital Content and Media Sciences Research Division, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan (e-mail: {mkastner, satoh}@nii.ac.jp)

²Graduate School of Informatics, Nagoya University, Chikusa-ku, Nagoya-shi, Aichi, 464-8601, Japan (e-mail: umemurak@cs.is.i.nagoya-u.ac.jp, {murase, ddeguchi}@nagoya-u.jp)

³Mathematical and Data Science Center, Nagoya University, Chikusa-ku, Nagoya-shi, Aichi, 464-8601, Japan (e-mail: ide@i.nagoya-u.ac.jp)

⁴Guardian Robot Project, Information R&D and Strategy Headquarters, RIKEN, Seika-cho, Kyoto, 619-0288, Japan (e-mail: yasutomo.kawanishi@riken.jp)

⁵Faculty of Human Environment, University of Human Environments, Okazaki-shi, Aichi, 444-3505, Japan (e-mail: t-hirayama@uhe.ac.jp)

⁶School of Engineering, Chukyo University, 101 Tokodachi, Kaizu-cho, Toyota-shi, Aichi, 470-0393, Japan (e-mail: kdoman@sist.chukyo-u.ac.jp)

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

Parts of this research were supported by JSPS KAKENHI 16H02846 program and Microsoft CORE-16 research program, and is a fruit of a joint-research project between National Institute of Informatics and Nagoya University.

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 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 *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

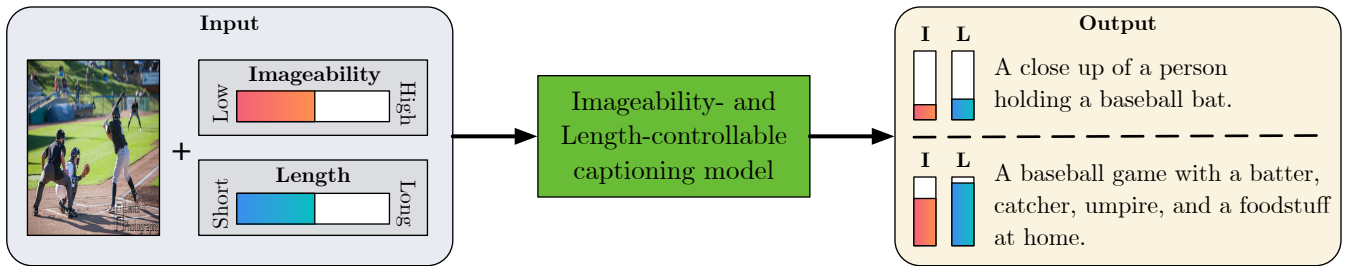


FIGURE 1. Proposed imageability- and length-controllable image captioning model. The *imageability* parameter allows for adjusting the visual descriptiveness on captions with the same length, while the *length* parameter changes the length for a fixed degree of visual descriptiveness. Both parameters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

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 for 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.

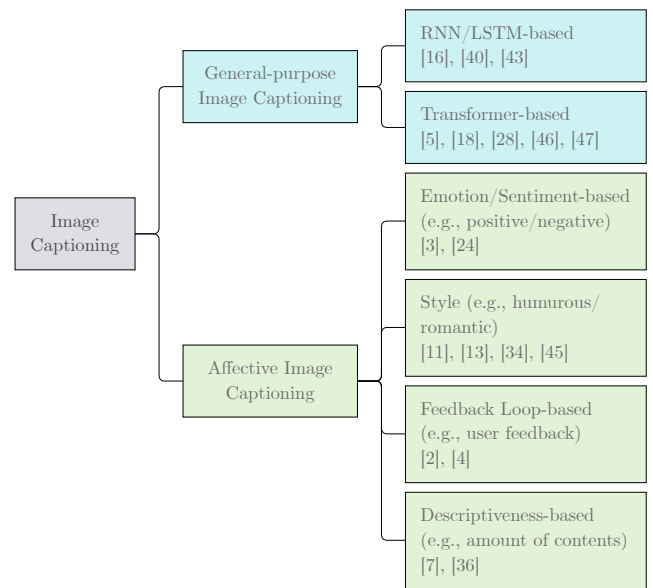


FIGURE 2. Related Work in Image Captioning. The related work is split into general-purpose and affective image captioning. The former tries to simply summarize image contents in a neutral short phrase, while the latter puts a strong focus on the emotion/sentiment, style, feedback, descriptiveness, or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

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.

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.

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.

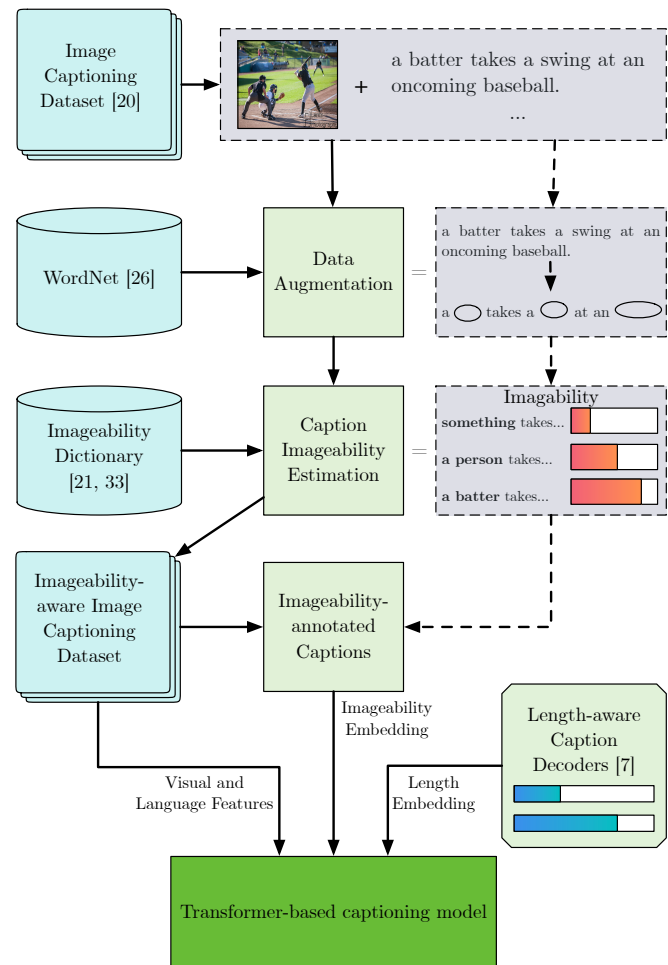


FIGURE 3. Flowchart of the proposed framework. A general purpose image captioning dataset is augmented using word substitutions through WordNet. This generates a diverse caption-dataset with different levels of visual descriptiveness. For each caption, an imageability score is calculated, which is then 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 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.

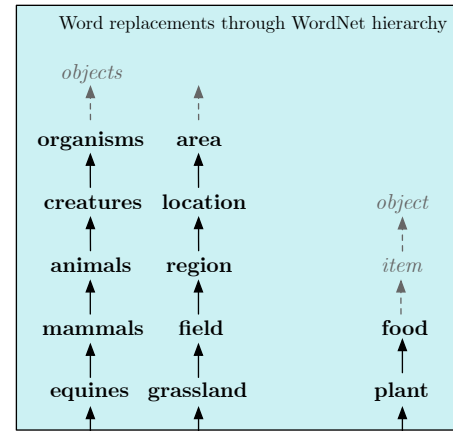
A. DATASET PREPARATION

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. 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.



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.

2) Caption Imageability Estimation

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

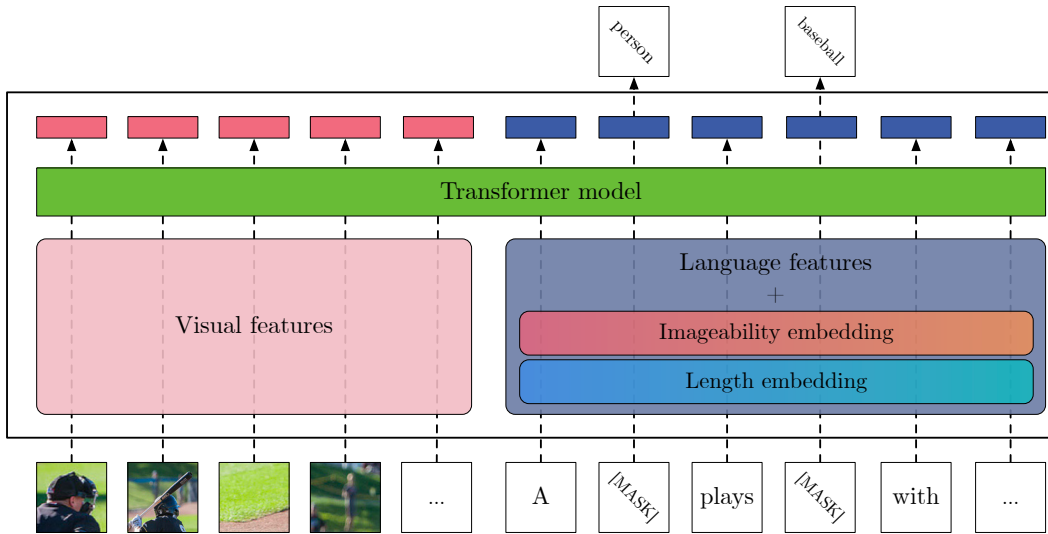


FIGURE 5. Proposed captioning model. The proposed model uses a transformer-based architecture. It is based on [7] which allows for length-controllable captioning. Inspired by their architecture, the proposed methods adds an imageability embedding layer which encodes the visual descriptiveness of captions.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

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

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 $C = \{c_i\}_{i=1}^N$, 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 pretrained on the Visual Genome dataset [17] to extract visual features. Using this object detection model, M objects in the regions $R = \{r_i\}_{i=1}^M$ 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

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

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 $C = \{c_i\}_{i=1}^N$ with c_i representing each word in a caption, 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 $c_i = [\text{MASK}]$ 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

a: 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 enough sufficiently available 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.

b: 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

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] can not 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
	Tell As You Imagine [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 method	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.
	Tell As You Imagine [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 method	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.
	Tell As You Imagine [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 method	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.

the imageability-controllable part of the transformer model is a contribution of this paper.

c: 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 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-

TABLE 3. Evaluation through general-purpose image captioning metrics. The proposed method is compared to [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	

controllable captions. Thus, we can not 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.

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.

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.

TABLE 4. Quantitative evaluation of imageability-controllable captions. The proposed method is compared to [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

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

TABLE 5. Subjective evaluation of visual descriptiveness. The proposed method is compared to [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	Avg.	95% CI	ρ			
			-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

before, TAYI uses an LSTM-based architecture while our 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.

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 [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. 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 B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision —ECCV 2016, 14th European Conf., Amsterdam, The Netherlands, Oct. 11–14, 2016, Procs., Part V, volume 9909 of Lecture Notes in Computer Science, pages 382–398. Springer, Cham, Switzerland, Oct. 2016.
- [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. 2020

¹<https://www.mturk.com/>

- IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 9962–9971, Seattle, WA, USA, June 2020.
- [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 V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision —ECCV 2018, 15th European Conf., Munich, Germany, Sept. 8–14, 2018, Procs., Part X*, volume 11214 of *Lecture Notes in Computer Science*, pages 527–543. Springer, Cham, Switzerland, Sept. 2018.
- [4] M. Cornia, L. Baraldi, and R. Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 8307–8316, Long Beach, CA, USA, June 2019.
- [5] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara. Meshed-memory transformer for image captioning. In Proc. 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 10578–10587, Online, June 2020.
- [6] M. J. Cortese and A. Fugett. Imageability ratings for 3,000 monosyllabic words. *Behav. Res. Methods Instrum. Comput.*, 36(3):384–387, Aug. 2004.
- [7] C. Deng, N. Ding, M. Tan, and Q. Wu. Length-controllable image captioning. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision —ECCV 2020, 16th European Conf., Glasgow, UK, Aug. 23–28, 2020, Procs., Part XIII*, volume 12358 of *Lecture Notes in Computer Science*, pages 712–729. Springer, Cham, Switzerland, Nov. 2020.
- [8] M. Denkowski and A. Lavie. METEOR Universal: Language specific translation evaluation for any target language. In Proc. 9th Workshop on Statistical Machine Translation, pages 376–380, Baltimore, MD, USA, June 2014.
- [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. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 10695–10704, Long Beach, CA, USA, June 2019.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. 2019 Conf. of the North American Chapter of ACL: Human Language Technologies, volume 1, pages 4171–4186, Minneapolis, MN, USA, June 2019.
- [11] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. StyleNet: Generating attractive visual captions with styles. In Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, pages 3137–3146, Honolulu, HI, USA, June 2017.
- [12] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. on Natural Language Processing, pages 6112–6121, Hong Kong, China, Nov. 2019.
- [13] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu. MSCap: Multi-style image captioning with unpaired stylized text. In Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 4204–4213, Long Beach, CA, USA, June 2019.
- [14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.
- [15] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pages 4565–4574, Las Vegas, NV, USA, June 2016.
- [16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition, pages 3128–3137, Boston, MA, USA, June 2015.
- [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.*, 123(1):32–73, May 2017.
- [18] G. Li, L. Zhu, P. Liu, and Y. Yang. Entangled transformer for image captioning. In Proc. 17th IEEE Int. Conf. on Computer Vision, pages 8928–8937, Seoul, Korea, Oct.–Nov. 2019.
- [19] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In Proc. 2004 ACL Workshop on Text Summarization Branches Out, pages 74–81, Barcelona, Cataluña, Spain, July 2004.
- [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 D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision —ECCV 2014, 13th European Conf., Zurich, Switzerland, Sept. 6–12, 2014, Procs., Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, Cham, Switzerland, Sept. 2014.
- [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 on Representation Learning for NLP, pages 217–222, Melbourne, VIC, Australia, July 2018.
- [22] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In Proc. 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 7219–7228, Salt Lake City, UT, USA, June 2018.
- [23] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In Proc. 52nd Annual Meeting of ACL: System Demonstrations, pages 55–60, Baltimore, MD, USA, June 2014.
- [24] A. P. Mathews, L. Xie, and X. He. SentiCap: Generating image descriptions with sentiments. In 30th AAAI Conf. on Artificial Intelligence, volume 30 of *Procs. AAAI Conf. on Artificial Intelligence*, pages 3574–3580. AAAI Press, Palo Alto, CA, USA, Feb. 2016.
- [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. 2020 ACM Int. Conf. on Multimedia Retrieval, pages 306–310, Online, Oct. 2020.
- [26] G. A. Miller. WordNet: A lexical database for English. *Comm. ACM*, 38(11):39–41, Nov. 1995.
- [27] A. Paivio, J. C. Yuille, and S. A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psychol.*, 76(1, Part 2):1–25, Jan. 1968.
- [28] Y. Pan, T. Yao, Y. Li, and T. Mei. X-linear attention networks for image captioning. In Proc. 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 10971–10980, Online, June 2020.
- [29] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: A method for automatic evaluation of machine translation. In Proc. 40th Annual Meeting of the Association for Computer Linguistics, pages 311–318, Philadelphia, PA, USA, July 2002.
- [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. 15th IEEE Int. Conf. on Computer Vision, pages 2641–2649, Santiago, Chile, Dec. 2015.
- [31] J. Reilly and J. Kean. Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cogn. Sci.*, 31(1):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 C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conf. on Neural Information Processing Systems 2015, Dec. 7–12, 2015, Montreal, QC, Canada*, volume 28 of *NeurIPS Procs.*, pages 91–99. Curran Associates, Red Hook, NY, 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*, 51:1258–1270, June 2019.
- [34] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. In Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 12516–12526, Long Beach, CA, USA, June 2019.
- [35] L. L. Thurstone. The method of paired comparisons for social values. *J. Abnorm. Psychol.*, 21(4):384–400, Jan. 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 J. Lokoč, T. Skopal, K. Schoeffmann, V. Mezaris, X. Li, S. Vrochidis, and I. Patras, editors, *MultiMedia Modeling —27th Int. Conf., MMM 2021, Prague, Czech Republic, June 22–24, 2021, Procs., Part II*, volume 12573 of *Lecture Notes in Computer Science*, pages 62–73. Springer, Cham, Switzerland, Jan. 2021.
- [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 I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems 2017, Dec. 4–9, 2017, Long Beach, CA, USA*, volume 30 of *NeurIPS Procs.*, pages 5998–6008. Curran Associates, Red Hook, NY, Dec. 2017.
- [38] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition, pages 4566–4575, Boston, MA, USA, June 2015.

- [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 32nd AAAI Conf. on Artificial Intelligence, volume 32 of Procs. AAAI Conf. on Artificial Intelligence, pages 7371–7379. AAAI Press, Palo Alto, CA, USA, Feb. 2018.
- [40] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition, pages 3156–3164, Boston, MA, USA, June 2015.
- [41] Q. Wang and A. B. Chan. Describing like humans: On diversity in image captioning. In Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 4190–4198, Long Beach, CA, USA, June 2019.
- [42] M. Wilson. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behav. Res. Methods Instrum. Comput.*, 20(1):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 *Int. Conf. on Machine Learning*, 7–9 July 2015, Lille, France, volume 37 of Procs. of Machine Learning Research, pages 2048–2057. ML Research Press, July 2015.
- [44] M. Zhang, R. Hwa, and A. Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In Proc. 2018 British Machine Vision Conf., number 8, pages 1–14, Newcastle upon Tyne, England, UK, Sept. 2018.
- [45] W. Zhao, X. Wu, and X. Zhang. MemCap: Memorizing style knowledge for image captioning. In 34th AAAI Conf. on Artificial Intelligence, volume 34 of Procs. AAAI Conf. on Artificial Intelligence, pages 12984–12992. AAAI Press, Palo Alto, CA, USA, Apr. 2020.
- [46] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and VQA. In 34th AAAI Conf. on Artificial Intelligence, volume 34 of Procs. AAAI Conf. on Artificial Intelligence, pages 13041–13049. AAAI Press, Palo Alto, CA, USA, Feb. 2020.
- [47] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu. Captioning transformer with stacked attention modules. *Appl. Sci.*, 8(5):739–749, May 2018.



MARC A. KASTNER received his BSc and MSc in Computer Science from Braunschweig University of Technology in Braunschweig, Germany, in 2013 and 2016, respectively. He received his PhD in Informatics in 2020 at the Graduate School of Informatics of Nagoya University, Japan. Since 2020, he has been working as a postdoctoral researcher at the National Institute of Informatics, Japan.

His research focuses on the connection of the human with multimedia, covering vision and language and affective computing related tasks.

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



KAZUKI UMEMURA received his BEng and MS from Nagoya University, Japan, in 2018 and 2020, respectively.

His research interests are in computer vision and natural language processing, focusing on the use of psycholinguistic features in image captioning.



ICHIRO IDE (M'12–SM'21) received his BEng, MEng, and PhD 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. Since 2020, he has been a Professor there. He was a Visiting Associate Professor at National Institute of Informatics from 2004 to 2010, an Invited Professor at Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), France in 2005, 2006, and 2007, a Senior Visiting Researcher at ISLA, Instituut voor Informatica, Universiteit van Amsterdam from 2010 to 2011.

His research interests range from 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 IEEE, IEICE, and IPS Japan, and a member of ACM, JSAI, and ITE.



YASUTOMO KAWANISHI received his BEng and MEng degrees in Engineering and a PhD degree in Informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. He became a Post Doctoral Fellow at Kyoto University, Japan in 2012. He moved to Nagoya University, Japan as a Designated Assistant Professor in 2014. In 2015, he became an Assistant Professor there. Since 2021, he has been a team leader at Multimodal Data Recognition Research Team,

RIKEN Guardian Robot Project, Japan.

His main research interests are robot vision for environmental understanding and computer vision for human understanding, especially pedestrian detection, tracking, retrieval, and recognition.

Dr. Kawanishi received the best paper award from SPC2009, and Young Researcher Award from IEEE ITS Society Nagoya Chapter. He is a member of IEICE, IIEEJ, and IEEE.



TAKATSUGU HIRAYAMA received the M.E. and D.E. degrees in Engineering Science from Osaka University in 2002 and 2005, respectively. From 2005 to 2011, he had been a Research Assistant Professor at the Graduate School of Informatics, Kyoto University, Japan. In 2011, he moved to the Graduate School of Information Science, Nagoya University, Japan. He became an Assistant Professor in 2012, a Designated Associate Professor in 2014, there. In 2017, he became a Designated Associate Professor at the Institutes of Innovation for Future Society, Nagoya University, Japan. Since 2021, he has been a Professor at the University of Human Environments, Japan.

His research interests include computer vision (face recognition, visual attention modeling, action recognition) and human-computer interaction (multi-modal interaction design, internal state estimation, interaction dynamics analysis).

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



KEISUKE DOMAN received his B.S. degree from the Department of Electrical and Electronic Engineering, his M.S. degree and Ph.D. from the Graduate School of Information Science at Nagoya University, Japan in 2007, 2009, and 2012, respectively. In 2013, he became a Lecturer at the School of Information Science and Technology, Chukyo University, Japan. Since 2020, he has been an Associate Professor at the School of Engineering, Chukyo University, Japan.

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, IEEE, ACM, and INSTICC.



DAISUKE DEGUCHI received his BEng and MEng in Engineering and PhD in Information Science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He is currently an Associate Professor in the Graduate School of Informatics, Nagoya University, Japan.

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 (M'87–SM'00–F'06) received his BEng, MEng, and PhD degrees in Electrical Engineering from Nagoya University, Japan. In 1980 he joined the Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993 he was a visiting research scientist at Columbia University, USA. From 2003 to 2021 he was a Professor at Nagoya University, Japan. he is currently a Professor Emeritus and Designated Professor, there.

He was awarded the IEICE Shinohara Award in 1986, the Telecom System Award in 1992, the IEEE CVPR (Conference on Computer Vision and Pattern Recognition) Best Paper Award in 1994, the IPS Japan Yamashita Award in 1995, the IEEE ICRA (International Conference on Robotics and Automation) Best Video Award in 1996, the Takayanagi Memorial Award in 2001, the IEICE Achievement Award in 2002, and the Ministry Award from the Ministry of Education, Culture, Sports, Science and Technology in 2003.

Dr. Murase is a Fellow of IEEE, IEICE, and IPS Japan.



SHIN'ICHI SATOH 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 has been a Full Professor with the National Institute of Informatics, Tokyo, since 2004. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, USA, from 1995 to 1997.

His current research interests include image processing, video content analysis, and multimedia databases.

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

• • •