Vol.2020-CVIM-222 No.5
2020/5/14

On the Quantification of the Mental Image of Visual Concepts for Multi-modal Applications

Marc A. Kastner[1,a)] Ichiro Ide[2,3] Yasutomo Kawanishi[3]
Takatsugu Hirayama[4] Daisuke Deguchi[3] Hiroshi Murase[3]

Abstract: The semantic gap is the lack of coincidence between the information one can extract from data and its interpretation. It is an yet solved issue for multimedia applications like image captioning, where it is often challenging to select the best tting wording out of a group of candidates. To create a measurement for the perceived di erences between words, this doctoral research proposes the idea of analyzing crawled image data to gain a better semantic understanding of language and vision. Abstract words have a broad mental image due to them being less visually de ned, while concrete words with a rather narrow visual feature space are visually easier to grasp. The core goal of this research is to approximate this perceived abstractness of those words as a metric. The thesis proposes two methods, looking at both relative and absolute measurements. For the relative method, a data-driven approach is proposed, while the absolute measurements train a machine-learned model on existing data from Psycholinguistics. Each method is evaluated using crowd-sourced data and compared to related approaches. With this, the thesis presents methods to analyze the mental image of words from di erent angles, targeting a way to quantify the semantic gap between vision and language.

Keywords: Multimedia Modeling, Language and Vision, Computational Psycholinguistics, Visual Concept Semantics

1. Introduction

With the rise of multimedia applications including various modalities like text, images, and videos, the need for a better understanding of the semantic relationship between language and vision becomes more and more prominent. While the relationship between language and vision is part of various research, the perception for humans is often only of secondary consideration. For vision applications, the detection of concrete objects is a heavily researched eld, but processing abstract ideas or concepts still easily causes issues, producing unnatural or unexpected results. A transfer of abstract ideas through modalities, e.g. from text to image, is an even harder problem, usually requiring extra information, and also knowledge on human expectation. A clearer modeling of how humans perceive language can improve word choice problems in various multimedia applications.

Through this doctoral research[30], the semantic gap between di erent concepts and words regarding human perception was studied. The central idea is to quantify the mental image of concepts, in order to nd a measurable metric of the degree of the semantic gap between di erent concepts. The mental image in this context is considered as \the creation or re-creation of an experience generated from memorial information" [41]. Regarding a concept or word, the goal is to analyze whether its mental image is *clear*, e.g. it creates a rather concrete image and thus is easy to depict, or its mental image is *vague*, meaning that is is rather abstract, can mean various things, or is not clearly visually de ned. For each input word, the methods' goal would be to estimate a value quantifying this idea on a scale. This overall goal is illustrated in Fig. 1.

The research is based on the core assumption that the average mental image regarding words across society is re- ected in the images available through the Web and Social Media. As such, if we can gather a su cient number of images related to a word or concept, the visual feature space will converge towards the average mental image of the said word or concept. Following, an exhaustive analysis of such images available through the Web and Social Media would yield meaningful information for estimating a metric predicting the mental image of a concept. This core assumption is illustrated in Fig. 2.

The doctoral thesis proposes two methods looking at this quanti cation from two angles: First, a relative measurement looking at granular di erences of related concepts, and second, an absolute measurement looking at the gen-

1 National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
2 Mathematical & Data Science Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
3 Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
4 Institute of Innovation for Future Society, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
a) mkastner@nii.ac.jp

ⓒ 2020 Information Processing Society of Japan

cept based on its popularity in social media. The method of counting the visual cluster calculates a distinct score for every term, describing its visual variety. Using a crowd-sourced survey, a ground-truth for this purpose has been obtained. When comparing the proposed datasets with another, it shows that the correlation to ground truth highly depends on the used recon guration. Compared to the baseline corpus, the recon guration using the proposed method with a Google Image Search weighting improved the measurements by 192 percent, showing very promising results in terms of understanding the relationship between vision and language.

4. Absolute Visual Variety Estimation for Arbitrary Concepts

The human perception of words has been researched as part of Psycholinguistics since the 1960s [49]. In this research, metrics called *Imageability* or *Concreteness* express the ability for a human to (visually) imagine a word. These concepts are commonly described on a Likert scale from harder-to-imagine words like *peaceful* to easier-to-imagine words like *cat*. A limited scale typically does not allow for detailed comparisons like in relative measurements, but the relationship between imageability and the previously discussed visual variety is indisputable. On a global dictionary-level scale, its estimation comes with its problems. While it is easy to compare the number of clusters between, e.g., *vehicle* and *taxi*, an evaluation of *peace* and *cat* has no evident reference point for comparison.

The second proposed method discussed in this doctoral research [29] targets an absolute estimation of visual variety for arbitrary concepts. As such, imageability as a metric closely ts this goal. However, previous research commonly obtains ground-truth annotations for imageability using crowd-sourced surveys or user studies. A method that estimates such labels has been researched using text data-mining [37]. However, a visual approach using images has not been part of research, despite the evident connection of visual information in the form of images and the ability to visually imagine something as a mental image. To x this gap, the proposed method analyzes the cross-similarity of Web-crawled images in order to train a model to estimate imageability as a metric for absolute visual variety.

In contrast to the work looking at relative visual variety estimation, this approach is mainly algorithm driven and does not assume a sophisticatedly created *ideal* dataset. This decision was initially out of convenience, as a data-driven method comes with its own bottlenecks through data acquisition processes and the amount of available data. Note that this second research method also targets an absolute measurements for an arbitrary domain. As such, there can not necessarily be a common WordNet hierarchy be assumed, as, e.g., *car* and *peace* do not share a meaningful hierarchy. As such, a bottom-up approach would not easily be applicable.

4.1 Imageability estimation

For a large number of images, a large set of related images is assumed. Using a selection of low- and high-level visual features is used to analyze the visual feature space of this concept. The visual features used for the experiments are discussed in Sec. 4.2. For each image and visual feature, a feature histogram is computed. Calculating the visual similarity between all images of the same concept, a similarity matrix is computed for each visual feature. The core assumption is similar to relative visual variety: A set of images for a high-imageability word would have somewhat similar images (e.g., all *cats* look similar), while a set of images for a low-imageability word is much more scattered (e.g., the concept *peace* is hard to visually de ne). From the obtained cross-similarity matrix, eigenvalues are computed. These eigenvalues intrinsically contain the trend of how visually similar a dataset is as individually encoded by each visual feature. The eigenvalues are used to train a regression model to predict the imageability of words. The proposed algorithm is outlined in Fig. 8.

4.2 Visual feature selection

In order to analyze each concept from multiple angles, a selection of features was chosen which look at visual characteristics from various viewpoints. Six visual features were chosen, which can logically be split into low-level and high-level features.

4.2.1 Low-level features

For lower-level features, the proposed method looks at traditional features used for CV pattern recognition.

First, the color distribution in the form of RGB color histograms was chosen to describe the overall mood of the images. In the context of imageability, this could also encode the atmosphere of an image, depending on e.g., colder or warmer colors. It also encodes information of visual adjectives like *yellow* or *bright*.

Second, global gradient descriptors in the form of GIST was chosen in order to describe global pattern distributions. Such features are commonly used for Web-retrieval engines.

Third, a local gradient descriptor in the form of Bag-of-Visual-Words (BoVW) [12] on SURF descriptorss [2] was chosen as a complementary feature. This type of feature is commonly used in traditional object detection, as it can decode which kind of patterns an object or scenes typically consist of.

4.2.2 High-level features

For high-level features, the proposed method uses pre-trained models, usually using deep learning, which approach image detection from a more *human* point of view. As such, rather than detecting raw edges, they are trained to encode human-annotated labels.

First, the theme of an image feature was chosen to encode the category of an image. This pre-trained model, rather than nding actual objects in an image, classi es an image into a category like *indoor*, *landscape*, or *architecture*. The setting of an image plays a large role in the similarity of

ⓒ 2020 Information Processing Society of Japan

Fig. 8: Flowchart of the proposed method for absolute visual variety estimation. Using a large number of images for a concept, the visual feature space is analyzed using a variety of low- and high-level visual features. Cross-similarity of all images is encoded in a similarity matrix, whose eigenvalues give indication of the general trend of visual similarity across the dataset. The features obtained are used to train a regression model with an existing imageability dictionary as ground-truth.

images, as it is largely an encoding of backgrounds, which are often the largest part of each image in terms of surface area. Concretely, a pre-annotated annotation part of the YFCC100M dataset [66] consisting of 1,570 classes was employed.

Second, the contents of an image feature was chosen to further describe an image based on what it contains. Here, the object detection model YOLO [52] is used to detect all objects in each image. From this, a distribution histogram is computed which is used as a feature.

Third, the composition of image feature was chosen to encode where objects appear in images of a certain concept. Here, again, the object detection model YOLO is used, but rather than creating a distribution histogram, the spatial coordinates of detected objects are mapped to a 10 10 grid. The resulting feature histograms describe the location of objects found in each image, based on this grid.

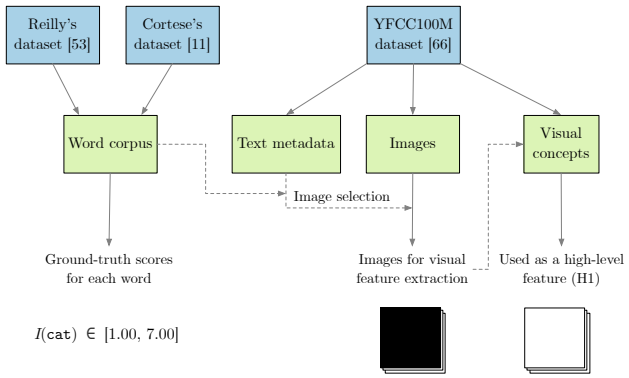

Fig. 9: Datasets used for this research. As imageability ground-truth annotations, two corpora from Psycholinguistics research are combined. The YFCC100M dataset is crawled for images and their text-annotations to obtain images for each concept.

4.3 Dataset construction

In this research, two types of datasets are employed. The rst is a dictionary with (English language) words and their imageability annotations, which provides the ground truth for both the training process and the evaluation. The second is a large number of images for each word, which will be used for visual feature extraction. The relationship between both types of datasets is illustrated in Fig. 9.

4.3.1 Imageability dictionary

There are a number of imageability dictionaries across various languages, including English [11][53], Indonesian [59], and Cantonese [72]. As described before, imageability dictionaries try to quantify the human perception of words, usually averaging the results over all test subjects. Low-imageability words would be things where one can not, or only hardly, grasp a mental image to describe it. These are usually rather abstract concepts, like *peace* or the word *abstract* itself. High-imageability words, on the other hand, are something rather concrete, usually easier to visually grasp, like *dog* or the color *red*.

Due to their nature, such dictionaries are commonly created by hand. Using crowd-sourcing or surveys, a set of words is judged by each test subject. Due to the amount of labor involved in this process, most dictionaries are rather small.

For this research, English imageability dictionaries by Reilly et al. [53] and Cortese et al. [11] in combination are used as a ground-truth for imageability. These datasets provide the results as a Likert scale score averaged over all test subjects, in the range of [1.0, 7.0]. While the former is only composed of nouns, the latter includes other parts-of-speech.

ⓒ 2020 Information Processing Society of Japan