International Journal of Multimedia Information Retrieval (2020) 9:31–45
https://doi.org/10.1007/s13735-019-00187-6

REGULAR PAPER

Characterization and classification of semantic image-text relations

Christian Otto[1] · Matthias Springstein[1] · Avishek Anand[2] · Ralph Ewerth[1,2]

Received: 10 September 2019 / Revised: 31 October 2019 / Accepted: 6 December 2019 / Published online: 22 January 2020

Abstract
The beneficial, complementary nature of visual and textual information to convey information is widely known, for example, in entertainment, news, advertisements, science, or education. While the complex interplay of image and text to form semantic meaning has been thoroughly studied in linguistics and communication sciences for several decades, computer vision and multimedia research remained on the surface of the problem more or less. An exception is previous work that introduced the two metrics *Cross-Modal Mutual Information* and *Semantic Correlation* in order to model complex image-text relations. In this paper, we motivate the necessity of an additional metric called *Status* in order to cover complex image-text relations more completely. This set of metrics enables us to derive a novel categorization of eight semantic image-text classes based on three dimensions. In addition, we demonstrate how to automatically gather and augment a dataset for these classes from the Web. Further, we present a deep learning system to automatically predict either of the three metrics, as well as a system to directly predict the eight image-text classes. Experimental results show the feasibility of the approach, whereby the predict-all approach outperforms the cascaded approach of the metric classifiers.

Keywords Image-text class · Multimodality · Data augmentation · Semantic gap

1 Introduction

In our digitized world, we are faced with multimodal information on a daily basis in various situations: consumption of news, entertainment, everyday learning or learning in formal education, social media, advertisements, etc. Different modalities help to convey information in an optimal manner, that is facilitating effective and efficient communication. For instance, please imagine to describe the exact shape of a leaf in textual form or, on the contrary, a specific date such as a birthday in solely visual[1] form. Neither of them is possible in a straightforward and comprehensible way and in general, it is not possible to translate every kind of information from one modality to another one. Although a quote says that "a picture is worth a thousand words," it is normally very difficult or even impossible to denote these thousand words. Thus, to appropriately make use of a single modality or two modalities is a key element for effective and efficient communication.

In a similar context, bridging the *semantic gap* has been identified as one of the key challenges in image retrieval (and multimedia) research [44], defined as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation." One challenge at this point in time was that information extraction from images was limited to low-level features. As a consequence, most multimedia and computer vision approaches aimed to solve the (perceptual) problem of object and scene recognition, considering visual concepts as semantic, high-level features. In fact, impressive progress has been reported for tasks such as object and visual concept recognition [12,23], or image captioning [1,20] in recent years. However, these approaches

✉ Christian Otto
christian.otto@tib.eu

Ralph Ewerth
ralph.ewerth@tib.eu

[1] TIB–Leibniz Information Centre for Science and Technology, Hannover, Germany

[2] L3S Research Center, Leibniz Universität Hannover, Hannover, Germany

[1] We use the term visual to refer to non-textual, pictorial information in this paper.

Fig. 1 An example of a complex message portrayed by an image-text pair elucidating the gap between the textual information and the image content. (*Source*: [17])

mostly address *only one possible interpretation* of visual content focusing on objects, persons, etc., but lack capabilities of human scene interpretation *going beyond the visible scene content*, i.e., interpreting symbols, gestures, and other contextual information.

Unfortunately, the complexity increases when we consider *multimodal information or cross-modal references* instead of *solely visual information*. The semantic gap is often caused (or enlarged) by a *modality gap*, since there is no direct translation between different modalities in general, as outlined above. In this work, we focus on the interplay of visual and textual information. An example is depicted in Fig. 1, which illustrates the interplay of interdependent textual and visual information. Today's state-of-the-art approaches normally do not contribute to answer intricate questions like "How much context or meaning is shared between text and image independent of the amount of shared concepts?" or "Does the type of information (or *image-text class*) match the current user query or retrieval scenario?". To answer such questions, a deeper understanding of the multimodal interplay of image and text and the resulting message is necessary. A challenge is that textual and related visual information are often not directly aligned; moreover, their interplay is typically complex and there is a large number of roles image and text can take on. In communication sciences and linguistics, this fact is often denoted as the "visual/verbal divide", which, for example, is well observable in comics or audiovisual data.

Recently, this research topic has gained some attention from some computer science researchers, who, either intentionally or unintentionally, assimilated ideas from communication sciences. Zhang et al. [53] investigate image-text relations in advertisements and distinguish between equivalent and non-equivalent parallel information transfer. They propose a method that automatically detects if the ad's slogan and pictorial component convey the same message independently, or if there is a bigger, mutual message. While this distinction is useful, it has been actually proposed before but was termed differently (e.g., *additive* and *parallel* [21], *independent* and *complementary* [30], and in a more general manner in own previous work [13,14]). Kruk et al. [24] tailor Marsh and White's [29] taxonomy to measure the author's intent of Instagram posts and two kinds of image-text relations, namely the *contextual relation* between the literal meanings of the image and caption, and the *semiotic relationship* between the meanings of the image and caption. To address Instagram posts, they suggest some additions to existing definitions, thus making their system less generalizable to other domains. In previous work, we have presented a more general approach [13,14] by introducing two metrics to describe image-text relations: cross-modal mutual information (CMI) and semantic correlation (SC). The metrics are based on the assumptions that visual and textual information can relate to each other a) based on their depicted or mentioned content, or b) based on their semantic context.

In this paper, we follow this paradigm and present the following contributions: First, we extend this set of two metrics by introducing a third metric called "Status," which is based on insights from linguistics and communication sciences. Second, we show how this set of metrics can be used to derive a set of eight semantic image-text classes, which are also coherent with studies and taxonomies from linguistics and communication science. Third, we demonstrate how to automatically gather samples from various Web resources in order to create a large (training) dataset, which we make publicly available. Finally, we present two baselines in form of deep learning systems to predict either the three metrics or directly the eight image-text classes. Compared to our conference paper at 2019 ACM International Conference on Multimedia Retrieval [37], this paper has been modified and extended as follows: Abstract, Introduction, and Conclusions are revised. The related work section is restructured and updated. The experimental evaluation is complemented with additional results and includes a comparison with our previous approach. Finally, an in-depth discussion of results is provided.

The remainder of the paper is organized as follows. Related work is discussed in Sect. 2. Section 3.1 introduces the third metric *Status* and provides definitions for all three metrics, while eight semantic image-classes are derived using these metrics in Sect. 3.2. Two deep learning baseline systems to predict either image-text metrics or semantic image-text classes are described in Sect. 4. Experiments are presented in Sect. 5, while Sect. 6 summarizes the paper and outlines areas for future work.

2 Related work

2.1 Multimedia information retrieval

Numerous publications in recent years deal with multimodal information in retrieval tasks. The general problem of reduc-

ing or bridging the semantic gap [44] between images and text is the main issue in cross-media retrieval [3,34,35,39,50]. Fan et al. [8] tackle this problem by modeling humans' visual and descriptive senses with a multi-sensory fusion network. They handle the *cognitive and semantic gap* by improving the comparability of heterogeneous media features and obtain good results for image-to-text and text-to-image retrieval. Liang et al. [26] propose a self-paced cross-modal subspace matching method by constructing a multimodal graph that preserves both the intra-modality and inter-modality similarity. Another application is targeted by Mazloom et al. [31], who extract a set of engagement parameters to predict the popularity of social media posts. While the confidence in predicting basic emotions like happiness or sadness can be improved by multimodal features [49], even more complex semantic concepts like sarcasm [42] or metaphors [43] can be predicted. This is enabled by evaluating the textual cues in the context of the image, providing a new level of semantic richness. The attention-based text embeddings introduced by Bahdanau et al. [2] analyze textual information under the consideration of previously generated image embeddings and improve tasks like document classification [51] and image caption generation [1,19,25].

A prerequisite to use heterogeneous modalities is the encoding in a joint feature space, which depends on the type of modality to encode, the number of training samples available, the type of classification to perform and the desired interpretability of the models [4]. One type of algorithms utilizes *Multiple Kernel Learning* [7,9]. Application areas are multimodal affect recognition [18,38], event detection [52], and Alzheimer's disease classification [28]. Deep neural networks can also be utilized to model multimodal embeddings. For instance, these systems can be used for the generation of image captions [20]; Ramanishka et al. [40] exploit audiovisual data and metadata, i.e., a video's domain, to generate coherent video descriptions "in the wild," using convolutional neural networks (CNN, ResNet [12]) to encode visual data. Alternative network architectures are GoogleNet [45] or DenseNet [15].

2.2 Communication sciences

The interpretation of multimodal information and the "visual/verbal divide" have been investigated in the field of visual communication and applied linguistics for many years.

One direction of research in recent decades has dealt with the assignment of image-text pairs to distinct image-text classes. In a pioneering work, Barthes [5] discusses the respective roles and functions of text and images. He proposes a first taxonomy, which introduces different types of (hierarchical) status relations between the modalities. If status is unequal, the classes *Illustration* and *Anchorage* are distinguished, otherwise their relation is denoted as *Relay*.

Martinec and Salway [30] extend Barthes' taxonomy and further divide the image-text pairs of *equal* rank into a *Complementary* and *Independent* class, indicating that the information content is either intertwined or equivalent in both modalities. They combine it with Halliday's [11] logico-semantics relations, which originally have been developed to distinguish text clauses. Martinec and Salway revised these grammatical categories to capture the specific logical relationships between text and image regardless of their *status*. McCloud [32] focuses on comic books, whose characteristic is that image and text typically do not share information by means of depicted or mentioned concepts, albeit they have a strong semantic connection. McCloud denotes this category as *Interdependent* and argues that "pictures and words go hand in hand to convey an idea that neither could convey alone." Other authors mention the case of negative correlations between the mentioned or visually depicted concepts (for instance, Nöth [36] or van Leeuwen [48]), denoting them *Contradiction* or *Contrast*, respectively. Van Leeuwen states that they can be used intentionally, e.g., in magazine advertisements by choosing opposite colors or other formal features to draw attention to certain objects.

2.3 Computable image-text relations

Henning and Ewerth [13,14] propose two metrics to characterize image-text relations in a general manner: *cross-modal mutual information* and *semantic correlation*. They suggest an autoencoder with multimodal embeddings to learn these relations while minimizing the need for annotated training data. Zhang et al. [53] investigate image-text relations in advertisements and distinguish, for instance, between equivalent parallel and non-equivalent parallel information transfer. However, they disregard previous work, e.g., in the field of communication science, and instead of using existing definitions (see next subsection) define their own set of relations. Kruk et al. [24] utilize Marsh and White's [29] taxonomy to model the author's intent of Instagram posts. Two kinds of image-text relations are suggested: the *contextual relation* between the literal meanings of the image and caption, and the *semiotic relationship* between the image and the caption.

3 Semantic image-text relations

The discussion of related work reveals that the complex cross-modal interplay of image and text has not been systematically modeled and investigated yet from a computer science perspective. In this section, we derive a categorization of classes of semantic image-text relations which can be used for multimedia information retrieval and Web search. This categorization is based on previous work in the fields of visual communication (sciences) and information retrieval.

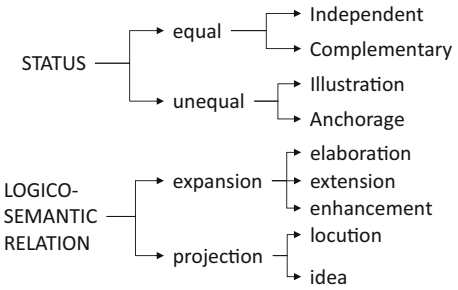

Fig. 2 Part of Martinec and Salway's taxonomy [30] that distinguishes image-text relation based on status (simplified)

However, one drawback of taxonomies in communication sciences is that their level of detail makes it sometimes difficult to assign image-text pairs to a particular class, as criticized by Bateman [6].

First, we evaluate the image-text classes described in communication science literature. As a point of departure, we consider Martinec and Salway's taxonomy (Fig. 2), which yields the classes *Illustration*, *Anchorage*, *Complementary*, and *Independent*. We disregard the class *Independent* since it is very uncommon that both modalities describe exactly the same information. Next, we introduce the class *Interdependent* suggested by McCloud [32], which in contrast to *Complementary* consists of image-text pairs where the intended meaning cannot be gathered from either of them exclusively. While a number of categorizations do not consider negative semantic correlations at all, Nöth [36], van Leeuwen [48], and Henning and Ewerth [13] consider this aspect. We believe that it is important for information retrieval tasks to consider negative correlations as well, for instance, in order to identify less useful multimodal information, contradictions, mistakes, etc. Consequently, we introduce the classes *Contrasting*, *Bad Illustration*, and *Bad Anchorage*, which are the negative counterparts for *Complementary*, *Illustration*, and *Anchorage*. Finally, we consider the case when text and image are *uncorrelated*.

While one objective of our work is to derive meaningful, distinctive, and comprehensible image-text classes, another contribution is their systematic characterization. For this purpose, we leverage the metrics cross-modal mutual information (CMI) and semantic correlation (SC) [13]. However, these two metrics are not sufficient to model a wide range of image-text classes. It is apparent that the *status* relation, originally introduced by Barthes [5], is adopted by the majority of taxonomies established in the last four decades (e.g., [30,47]), implying that this relation is essential to describe an image-text pair. It portrays how two modalities can relate to one another in a hierarchical way reflecting their relative importance. Either the text supports the image (*Anchorage*), or the image supports the text (*Illustration*), or both modalities contribute equally to the overall meaning (e.g., *Complementary*. This encourages us to extend the two-dimensional feature space of CMI and SC with the *status* dimension (STAT). In the next section, we provide definitions for the three metrics and subsequently infer a categorization of semantic image-text classes from them. Our goal is to reformulate and clarify the interrelations between visual and textual content in order to make them applicable for multimodal indexing and retrieval. An overview of the image-text classes and their mapping to the metrics, as well as possible use cases is given in Fig. 3.

3.1 Metrics for image-text relations

Concepts and entities The following definitions are related to concepts and entities in images and text. Generally, plenty of concepts and entities can be found in images ranging from the main focus of interest (e.g., a person, a certain object, an event, a diagram) to barely visible or background details (e.g., a leaf of grass, a bird in the sky). Normally, the meaning of an image is related to the main objects in the foreground. When assessing relevant information in images, it is reasonable to regard these concepts and entities, which, however, adds a certain level of subjectivity in some cases. But most of the time the important entities can be easily determined.

Cross-modal mutual information (CMI) Depending on the (fraction of) mutual presence of concepts and entities in both image and text, the cross-modal mutual information ranges from 0 (no overlap of depicted concepts) to 1 (concepts in image and text overlap entirely).

	Uncorrelated CMI=0 SC=0 STAT=0	Interdependent CMI=0 SC=1 STAT=0	Complementary CMI=1 SC=1 STAT=0	Illustration CMI=1 SC=1 STAT=T	Anchorage CMI=1 SC=1 STAT=I	Constrasting CMI=1 SC=-1 STAT=0	Bad Illustration CMI=1 SC=-1 STAT=T	Bad Anchorage CMI=1 SC=-1 STAT=I
What is captured?	No shared concepts or semantic backgrounds	No shared concepts, but joint message on a higher semantic level	Modalities complement each other	Text is supplemented with an exchangable image	Image is supplemented with a caption describing visual concepts	Modalities complement each other, but contain contrasting details	Given visual example is ill composed, unusual of ambiguous	A given caption describes details of displayed information incorrectly
Possible Usecases	• Filter for retrieval tasks • Adblocker	• Adblocker • Marketing retrieval tasks	• Recommender systems • Cross-modal retrieval • Web search	• Search tasks in educational settings • Text books	Search tasks in educational settings, e.g., definitions or explanations	• Quality check • Filter for retrieval tasks or recommender systems	• Quality check • Filter for retrieval tasks or recommender systems	• Quality check • Filter for retrieval tasks or recommender systems

Fig. 3 Overview of the proposed image-text classes and their potential use cases

It is important to point out that CMI ignores a deeper semantic meaning, in contrast to *semantic correlation*. If, for example, a small man with a blue shirt is shown in the image, while the text talks about a tall man with a red sweater, the CMI would still be positive due to the mutual concept "man." But since the description is confusing and hinders interpretation of the multimodal information, semantic correlation (SC, see below) of this image-text pair would be negative. Image-text pairs with high CMI can be found in image captioning datasets, for instance. The images and their corresponding captions have a descriptive nature, which is why they have explicit representations in both modalities. In contrast, news articles or advertisements often have a loose connection to their associated images by means of mutual entities or concepts. The range of cross-modal mutual information (CMI) is [0, 1].

Semantic correlation (SC) The (intended) meaning of image and text can range from coherent (SC $=$ 1), over uncorrelated (SC $=$ 0) to contradictory (SC $=-1$). This refers to concepts, descriptions and interpretation of symbols, metaphors, as well as to their relations to one another. Typically, an interpretation requires contextual information, knowledge, or experience and it cannot be derived exclusively from the entities in the text and the objects depicted in the image. The range of possible values is $[-1, 1]$, where a negative value indicates that the co-occurrence of an image and a text is contradicting and disturbs the comprehension of the multimodal content. This is the case if a text refers to an object in an image and cannot be found there, or has different attributes as described in the text. An observer might notice a contradiction and ask herself "Do image and text belong together at all, or were they placed jointly by mistake?". A positive score on the contrary suggests that both modalities share a semantic context or meaning. The third possible option is that there is no semantic correlation between entities in the image and the text, yielding SC = 0.

Status (STAT) Status describes the hierarchical relation between an image and text with respect to their relative importance. Either the image is "subordinate to the text" (stat $= T$), implying an exchangeable image which plays the minor role in conveying the overall message of the image-text pair, or the text is "subordinate to the image" (stat = I), usually characterizing text with additional information (e.g., a caption) for an image that is the center of attention. An *equal status* (stat = 0) describes the situation where image and text are equally important to convey the overall message.

Images which are "subordinate to text" (class *Illustration*) "elucidate" or "realize" the text. This is the case, if a text describes a general concept and the associated image shows a concrete example of that concept. Examples for the class *Illustration* can be found in textbooks and encyclopedias. On the contrary, in the class *Anchorage* the text is "subordinate to the image." This is the case, if the text answers the question "What can be seen in this image?". It is common that direct references to objects in the image can be found and the readers are informed what they are looking at. This type of image-text pair can be found in newspapers or scientific documents, but also in image captioning datasets. The third possible state of a *status relation* is "equal," which describes an image-text pair where both modalities contribute individually to the conveyed information. Also, either part contains details that the other one does not. According to Barthes [5], this class describes the situation where the information depicted in either modality is part of a more general message and together they elucidate information on a higher level that neither could do alone.

3.2 Defining classes of image-text relations

In this section, we show how the combination of our three metrics can be naturally mapped to distinctive image-text classes (see also Fig. 3). For this purpose, we simplify the data value space for each dimension. The level of semantic correlation can be represented by the interval $[-1, 1]$. Henning and Ewerth [13,14] distinguish five levels of CMI and SC. In this work, we omit these intermediate levels since the general idea of positive, negative, and uncorrelated image-text pairs is sufficient for the task of assigning image-text pairs to distinct classes. Therefore, the possible states of semantic correlation (SC) are sc $\in \{-1, 0, 1\}$. For a similar reason, finer levels for CMI are omitted, resulting in two possible states for cmi $\in \{0, 1\}$, which correspond to *no overlap* and *overlap*. Possible states of status are stat $\in \{T, 0, I\}$: *image subordinate to text* (stat $= T$), *equal status* (stat = 0), and *text subordinate to image* (stat = I).

If approached naively, there are $2 \times 3 \times 3 = 18$ possible combinations of SC, CMI, and STAT. A closer inspection reveals that (only) eight of these classes match with existing taxonomies in communication sciences, confirming the coherence of our analysis. The remaining ten classes can be discarded since they cannot occur in practice or do not make sense. The reasoning is given after we have defined the eight classes that form the categorization (Fig. 4).

Uncorrelated (cmi = 0, sc = 0, stat = 0) This class contains image-text pairs that do not belong together in an obvious way. They neither share entities and concepts nor there is an interpretation for a semantic correlation (e.g., see Fig. 5, left).

Complementary (cmi = 1, sc = 1, stat = 0) The class *Complementary* comprises the classic interplay between visual and textual information, i.e., both modalities share information but also provide information that the other one does not. Neither of them is dependent on the other one and their status is equal. It is important to note that the amount of information is not necessarily the same in both modalities. The most significant factor is that an observer is still able to understand the

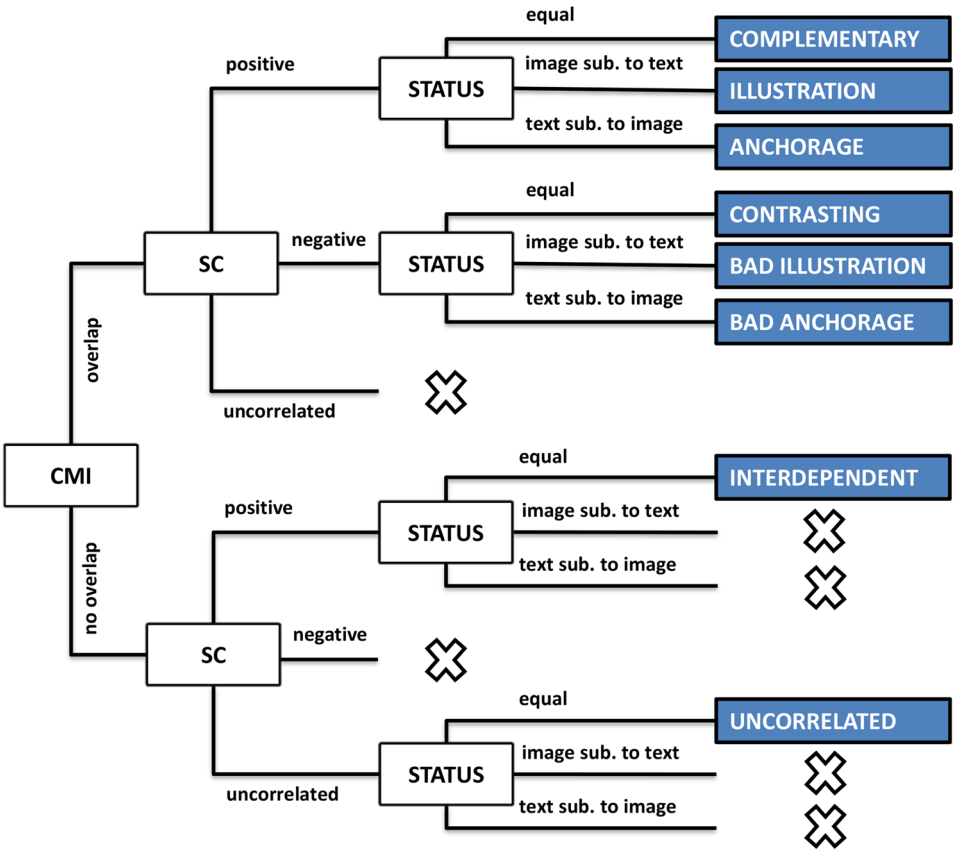

Fig. 4 Our categorization of image-text relations. Discarded subtrees or leaves are marked by an **X** for clarity. Please note that there are no hierarchical relations implied

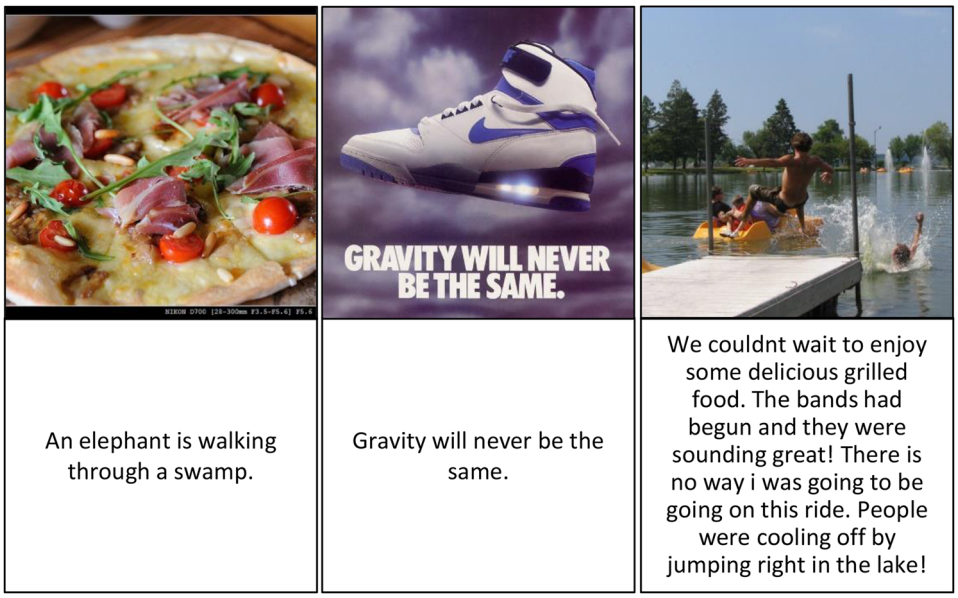

Fig. 5 Examples for the *Uncorrelated* (left), *Interdependent* (middle) and *Complementary* (right) classes. (*Sources*: see Sect. 4.1)

key information provided by either of the modalities alone (Fig. 5, right). The definitions of the next two classes will clarify that further.

Interdependent (cmi = 0, sc = 1, stat = 0) This class includes image-text pairs that do not share entities or concepts by means of mutual information, but are related by a semantic context. As a result, their combination conveys a new meaning or interpretation which neither of the modalities could have achieved on its own. Such image-text pairs are prevalent in advertisements where companies combine eye-catching images with funny slogans supported by metaphors or puns, without actually naming their product (Fig. 5, middle). Another genre that relies heavily on these *inter-*

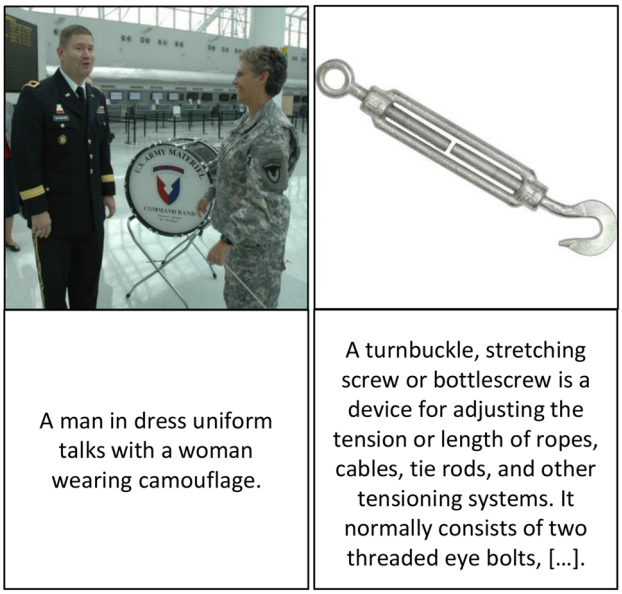

Fig. 6 Examples for the *Anchorage* (left) and *Illustration* (right) classes. (*Sources*: see Sect. 4.1)

dependent examples are comics or graphic novels, where speech bubbles and accompanying drawings are used to tell a story. Interdependent information is also prevalent in movies and TV material in the auditory and visual modalities.

Anchorage (cmi = 1, sc = 1, stat = I) On the contrary, the *Anchorage* class is an image description and acts as a supplement for an image. Barthes states that the role of the text in this class is to fix the interpretation of the visual information as intended by the author of the image-text pair [5]. It answers the question "What is it?" in a more or less detailed manner. This is often necessary since the possible meaning or interpretation of an image can noticeably vary and the caption is provided to pinpoint the author's intention. Therefore, an *Anchorage* can be a simple image caption, but also a longer text that elucidates the hidden meaning of a painting. It is similar to *Complementary*, but the main difference is that the text is subordinate to image in *Anchorage* (see Fig. 6).

Illustration (cmi = 1, sc = 1, stat = T) The class *Illustration* contains image-text pairs where the visual information is subordinate to the text and has therefore a lower *status*. An instance of this class could be, for example, a text that describes a general concept and the accompanying image depicts a specific example (Fig. 6). A distinctive feature of this class is that the image is replaceable by a very different image without rendering the constellation invalid. If the text is a definition of the term "mammal," it does not matter if the image shows an elephant, a mouse, or a dolphin. Each of these examples would be valid in this scenario. In general, the text is not dependent on the image to provide the intended information.

Contrasting (cmi = 1, sc = − 1, stat = 0)
Bad Illustration (cmi = 1, sc = − 1, stat = T)
Bad Anchorage (cmi = 1, sc = − 1, stat = I)

These three classes are the counterparts to *Complementary, Illustration*, and *Anchorage*: They share their primary features, but have a **negative SC** (see Fig. 7). In other words, the transfer of knowledge is impaired due to inconsistencies or contradictions when jointly viewing image and text [13]. In contrast to *uncorrelated* image-text pairs, these classes share information and obviously they belong together in a certain way, but particular details or characteristics are contradicting. For instance, a *Bad Illustration* pair could consist of a textual description of a bird, whose most prominent feature is its colorful plumage, but the bird in the image is actually a gray pigeon. This can be confusing and an observer might be unsure if she is looking at the right image. Similarly, contradicting textual counterparts exist for each of these classes. In Sect. 4.1, we describe how we generate training samples for these classes.

3.3 Impossible image-text relations

The eight classes described above form the categorization as shown in Fig. 4. The following ten combinations of metrics were discarded, since they do not yield meaningful image-text pairs.

Cases A cmi = 0, sc = − 1, stat = T, 0, I These three classes cannot exist: If the shared information is zero, then there is nothing that can contradict one another. As soon as a textual description relates to a visual concept in the image, there is cross-modal mutual information and CMI > 0.

Cases B cmi = 0, sc = 0, stat = T, I The metric combination cmi = 0, sc = 0, stat = 0 describes the class *Uncorrelated* of image-text pairs which are neither in contextual nor visual relation to one another. Since it is not intuitive that a text is subordinate to an uncorrelated image or vice versa, these two classes are discarded.

Cases C cmi = 0, sc = 1, stat = T, I Image-text pairs in the class *Interdependent* (cmi = 0, sc = 1, stat = 0) are characterized by the fact that even though they do not share any information they still complement each other by conveying additional or new meaning. Due to the nature of this class a subordination of one modality to the other one is not plausible: Neither of the conditions for the states *image subordinate to text* and *text subordinate to image* is fulfilled due to lack of shared concepts and entities. Therefore, these two classes are discarded.

Cases D cmi = 1, sc = 0, stat = T, 0, I As soon as there is an overlap of essential depicted concepts, there has to be a minimum of semantic overlap. We consider entities as essential, if they contribute to the overall information or meaning of

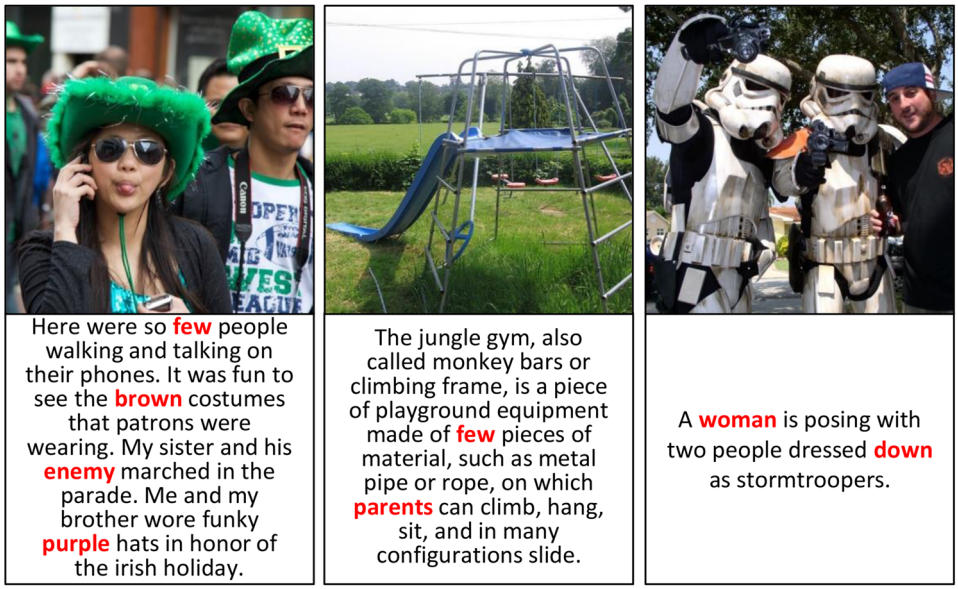

Fig. 7 Examples for the *Contrasting* (left), *Bad Illustration* (middle), and *Bad Anchorage* (right) classes. (*Sources*: see Sect. 4.1)

the image-text pair. This excludes trivial background information such as the type of hat a person wears in an audience behind a politician giving a speech. The semantic correlation can be minor, but it would still correspond to SC = 1 according to the definition above. Therefore, the combination cmi = 1, sc = 0 and the involved possible combinations of STAT are discarded.

4 Predicting image-text classes

In this section, we present our approach to automatically predict the introduced image-text metrics and classes. We propose a deep learning architecture that realizes a multimodal embedding for textual and pictorial data. Deep neural networks achieve better results, when they are trained with a large amount of data. However, for the addressed task no such dataset exists. Crowdsourcing is an alternative to avoid the time-consuming task of manually annotating training data on our own, but requires significant efforts to maintain the quality of annotations obtained in this way. Therefore, we follow two strategies to create a sufficiently large training set. First, we automatically collect image-text pairs from different open access Web sources. Second, we suggest a method for training data augmentation (Sect. 4.1) that allows us to also generate samples for the image-text classes that rarely occur on the Web, for instance, *Bad Illustration*. We suggest two classifiers, a **"classic"** approach, which simply outputs the most likely image-text class, as well as a cascaded approach based on classifiers for the three metrics. The motivation for the latter is to divide the problem into three easier classification tasks. Their subsequent **"cascaded"** execution will still lead us to the desired output of image-text classes according to Fig. 4. The deep learning architecture is explained in Sect. 4.2.

4.1 Training data augmentation

The objective is to acquire a large training dataset of high quality image-text pairs with a minimum effort in manual labor. On the one hand, there are classes like *Complementary* or *Anchorage* available from a multitude of sources and can therefore be easily crawled. Other classes like *Uncorrelated* do not naturally occur in the Web, but can be generated with little effort. On the other hand, there are rare classes like *Contrasting* or *Bad Anchorage*. While they do exist and it is desirable to detect these image-text pairs as well (see Fig. 3), there is no abundant source of such examples that could be used to train a robust classifier.

Only few datasets are publicly available that contain images and corresponding textual information, which are not simply based on tags and keywords but also use cohesive sentences. Two examples are the image captioning dataset MSCOCO [27] as well as the Visual Storytelling dataset VIST [16]. A large number of examples can be easily taken from these datasets, namely for the classes *Uncorrelated*, *Complementary*, and *Anchorage*. Specifically, the underlying hierarchy of MSCOCO is exploited to ensure that two randomly picked examples are not semantically related to one another and then join the caption of one sample with the image of the other one to form *Uncorrelated* samples. In this way, we gathered 60 000 ***uncorrelated*** training samples.

The VIST dataset has three types of captions for their five-image-stories. The first one "Desc-in-Isolation" resembles the generic image-caption dataset and can be used to generate examples for the class ***Anchorage***. These short descriptions

Table 1 Distribution of class labels in the generated dataset

Class	No. of samples
Uncorrelated	60,000
Interdependent	1007
Complementary	33,088
Illustration	5447
Anchorage	62,637
Contrasting	31,368
Bad Illustration	4099
Bad Anchorage	27,210

Table 2 Distribution of metric labels in the generated dataset

Class	No. of samples
STAT T	125,463
STAT 0	9546
STAT I	89,847
SC -1	62,677
SC 0	60,000
SC 1	102,179
CMI 0	61,007
CMI 1	163,849

are similar to MSCOCO captions, but slightly longer, so we decided to use them. Around 62 000 examples have been generated this way. The pairs represent this class well, since they include textual descriptions of the visually depicted concepts without any low-level visual concepts or added interpretations. More examples could have been generated similarly, but we have to restrict the level of class imbalance. The second type of VIST captions "Story-in-Sequence" is used to create ***Complementary*** samples by concatenating the five captions of a story and pairing them randomly with one of the images of the same story. Using this procedure, we generated 33 088 examples.

While there are certainly much more possible constellations of *complementary* content from a variety of sources, the various types of stories of this dataset give a solid basis. The same argumentation holds for the ***Interdependent*** class. Admittedly, we had to manually label a set of about 1 007 entries of Hussain et al.'s Internet Advertisements dataset [17] to generate these image-text pairs. While they exhibit the right type of image-text relations, the accompanied slogans (in the image) are not annotated separately and optical character recognition did not achieve high accuracy due to ornate fonts, etc. Furthermore, some image-text pairs had to be removed, since some slogans specifically mention the product name. This contradicts the condition that there is no overlap between depicted concepts and textual description, i.e., $cmi = 0$.

The ***Illustration*** class is established by combining one random image for each concept of the ImageNet dataset [41] with the summary of the corresponding article of the English Wikipedia, in case it exists. This nicely fits the nature of the class since the Wikipedia summary often provides a definition including a short overview of a concept. An image of the ImageNet class with the same name as the article should be a replaceable example image of that concept.

The three remaining classes ***Contrasting, Bad Illustration*** and ***Bad Anchorage*** occur rarely and are hard to detect automatically. Therefore, it is not possible to automatically crawl a sufficient amount of samples. To circumvent this problem, we suggest to transform the respective positive counterparts by replacing 530 keywords [37] (adjectives, directional words, colors) by antonyms and opposites in the textual description of the positive examples to make them less comprehensible. For instance, "tall man standing in front of a green car" is transformed into a "small woman standing behind a red car." While this does not absolutely break the semantic connection between image and text, it surely describes certain attributes incorrectly which impairs the accurate understanding and subsequently justifies the label of $sc = -1$. This strategy allows us to transform a substantial amount of the "positive" image-text pairs into their negative counterparts. Finally, for all classes we truncated the text if it exceeded 10 sentences. In total, the dataset consists of 224 856 image-text pairs. Tables 1 and 2 give an overview about the data distribution, first sorted by class and the second one according to the distribution of the three metrics, which were also used in our experiments.

4.2 Design of the deep classifiers

As mentioned above, we introduce two classification approaches: "classic" and "cascade." The advantage of the latter is that it is easier to maintain a better class balance of samples, while it is also the easier classification problem. For instance, example data of the classes *Contrasting*, *Bad Illustration*, and *Bad Anchorage* are used to train the neural network how negative semantic correlation looks like. This should make the training process more robust against overfitting and underfitting, but naturally also increases the training and evaluation time by a factor of three.

Both methods follow the architecture shown in Fig. 8, but for "cascade" three networks have to be trained and subsequently applied to predict an image-text class. To encode the input image, the deep residual network "Inception-ResNet-v2" [45] is used, which is pre-trained on the dataset of the ImageNet challenge [41]. To embed this model in our system, we remove all fully connected layers and extract the

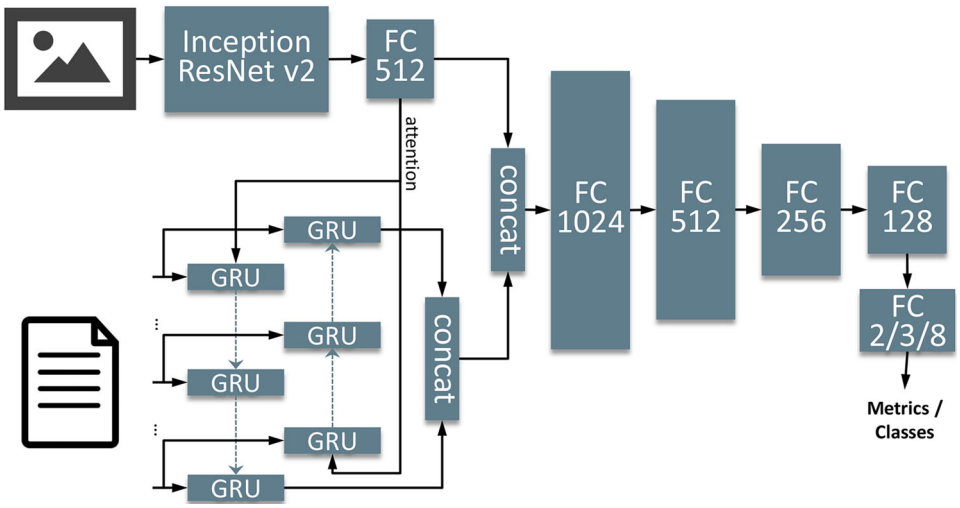

Fig. 8 General structure of the deep learning system with multimodal embedding. The last fully connected layer (FC) has 2, 3, or 8 outputs depending on whether CMI (two levels), SC/STAT (three levels), or all eight image-text classes ("classic" approach) are classified

feature maps with an embedding size of 2048 from the last convolutional layer.

The text is encoded by a pre-trained model of the word2vec [33] successor fastText [10], which has the remarkable ability to produce semantically rich feature vectors even for unknown words. This is due to its skip-gram technique, which does not observe words as a whole but as n-grams, that is a sum of word parts. For instance, the word *library* would be decomposed into the following tri-grams: ⟨ li, lib, ibr, bra, rar, ary, ry ⟩.

Thus, it enables the system to recognize a word or derived phrasings despite of typing errors. FastText utilizes an embedding size of 300 for each word and we feed them into a bidirectional GRU (gated recurrent unit) inspired by Yang et al. [51], which reads the sentence(s) forwards and backwards before subsequently concatenating the resulting feature vectors. In addition, an attention mechanism is incorporated through another convolutional layer, which reduces the image encoding to 300 dimensions, matching the dimensionality of the word representation set by fastText. In this way it is ensured that the neural network reads the textual information under the consideration of the visual features, which enforces it to interpret the features in unison. The final text embedding has a dimension of 1024. After concatenating image (to get a global feature representation from the image, we apply average pooling to the aforementioned last convolutional layer) and text features, four consecutive fully connected layers (dimensions: 1024, 512, 256, 128) comprise the classification layer. This layer has two outputs for CMI, three outputs for SC and STAT, or eight outputs for the "classic" classifier, respectively. For the actual classification process in the cascade approach, the resulting three models have to be applied sequentially in an arbitrary order. We select the order CMI ⇒ SC ⇒ STAT, the evaluations of the three classifiers yield the final assignment to one of the eight image-text classes (Fig. 4).

5 Experimental evaluation

The dataset was split into a training set and a manually verified test set to ensure high quality labels. It initially contained 800 image-text pairs, where for each of the eight classes 100 examples were taken out of the automatically crawled and augmented data. The remaining 239,307 examples were used to train the four different models (three for the "cascade" classifier and one for the "classic" approach) for 100,000 iterations each with the TensorFlow framework. The *Adam optimizer* was used with its standard learning rate and a dropout rate of 0.3 for the image embedding layer and 0.4 for the text embedding layer. Also a softmax cross entropy loss was used and a batch size of 12 on a NVIDIA Titan X. All images were rescaled to a size of 299 × 299 and Szegedy et al.'s [46] image preprocessing techniques were applied. This includes random cropping of the image as well as random brightness, saturation, hue and contrast distortion to avoid overfitting. In addition, we limit the length of the textual information to 50 words per sentence and 30 sentences per image-text pair. All "Inception-ResNet-v2" layers were pre-trained with the ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2010 [41] dataset to reduce the training effort. The training and test data are publicly available at https://doi.org/10.25835/0010577.

5.1 Experimental results

To assure highly accurate ground truth data for our test set, we asked three persons of our group (one of them is a co-author) to manually annotate the 800 image-text pairs.

Each annotator received an instruction document that contained short definitions of the three metrics (Sect. 3.1), the categorization in Fig. 4, and one example per image-text class (similar to Figs. 5, 6, 7). The inter-coder agreement has been evaluated using Krippendorff's alpha [22] and yielded a value

of $\alpha = 0.847$ (across all annotators, samples, and classes). A class label was assigned, if the majority of annotators agreed on it for a sample. Besides, the eight image-text classes, the annotators could also mark a sample as *Unsure* which denotes that an assignment was not possible. If *Unsure* was the majority of votes, the sample was not considered for the test set. This only applied for two pairs, which reduced the size of the final test set to 798.

Table 3 Comparison of the automatically generated labels with the annotations of the three volunteers (i.e., ground truth data) and the resulting number of samples per class in the test set

Class	Uncorr.	Interdep.	Compl.	Illustration
Recall	69.2%	97.6%	83.8%	83.7%
Precision	98.7%	96.3%	88.0%	80.7%
#Samples	149	100	106	95
Class	Anchorage	Contrasting	Bad Illu.	Bad Anch.
Recall	90.3%	89.0%	98.6%	91.9%
Precision	87.3%	78.3%	69.0%	87.0%
#Samples	95	87	71	95

Comparing the human labels with the automatically generated labels allowed us to evaluate the quality of the data acquisition process. Therefore, we computed how good the automatic labels matched with the human ground truth labels (Table 3). The low recall for the class *Uncorrelated* indicates that there were uncorrelated samples in the other data sources that we exploited. The *Bad Illustration* class has the lowest precision and was mostly confused with *Illustration* and *Uncorrelated*, that is the human annotators considered the automatically "augmented" samples either as still valid or uncorrelated.

The results for predicting image-text classes using both the "classic approach" (Table 5) and the "cascade approach"

Table 4 Confusion matrix for the "cascade" classifier on the test set of 798 image-text pairs

Class	Undef.	Uncorrelated	Interdep.	Compl.	Illustration	Anchorage	Contrasting	Bad Illust.	Bad Anch.	Sum
Undefined	**0**	0	0	0	0	0	0	0	0	0
Uncorrelated	2	**96**	4	7	21	1	4	13	1	149
Interdependent	3	3	**92**	1	0	1	0	0	0	100
Complementary	1	0	1	**93**	0	2	9	0	0	106
Illustration	1	0	0	0	**82**	0	0	12	0	95
Anchorage	11	4	5	25	1	**41**	2	1	5	95
Contrasting	0	0	0	2	0	0	**85**	0	0	87
Bad Illustration	0	0	0	0	8	0	0	**63**	0	71
Bad Anchorage	9	2	0	4	0	6	33	0	**41**	95
Precision	–	91.43%	90.20%	70.45%	73.21%	80.39%	63.91%	70.79%	87.23%	–
Recall	–	64.43%	92.00%	87.74%	86.32%	43.16%	97.70%	88.73%	43.16%	–

Bold values indicate the number of correctly classified samples on the main diagonal of the confusion matrix
The rows show the ground truth, while the columns show the predicted samples

Table 5 Confusion matrix for the "classic" classifier on the test set of 798 image-text pairs

Class	Undef.	Uncorrelated	Interdep.	Compl.	Illustration	Anchorage	Contrasting	Bad Illust.	Bad Anch.	Sum
Uncorrelated	–	**67**	3	5	23	34	5	11	1	149
Interdependent	–	0	**94**	0	0	5	0	0	1	100
Complementary	–	0	0	**93**	0	4	9	0	0	106
Illustration	–	0	0	0	**84**	0	0	11	0	95
Anchorage	–	2	2	0	2	**83**	0	0	6	95
Contrasting	–	0	0	3	0	0	**84**	0	0	87
Bad illustration	–	0	0	0	2	0	0	**69**	0	71
Bad anchorage	–	2	0	0	0	21	1	0	**71**	95
Precision	–	94.4%	94.9%	92.1%	75.7%	56.5%	84.8%	75.8%	89.9%	–
Recall	–	45.0%	94.0%	87.7%	88.4%	87.4%	96.5%	97.2%	74.7%	–

Bold values indicate the number of correctly classified samples on the main diagonal of the confusion matrix
The rows show the ground truth, while the coloumns show the predicted samples. (Undefined column was added for better comparability with Table 4)

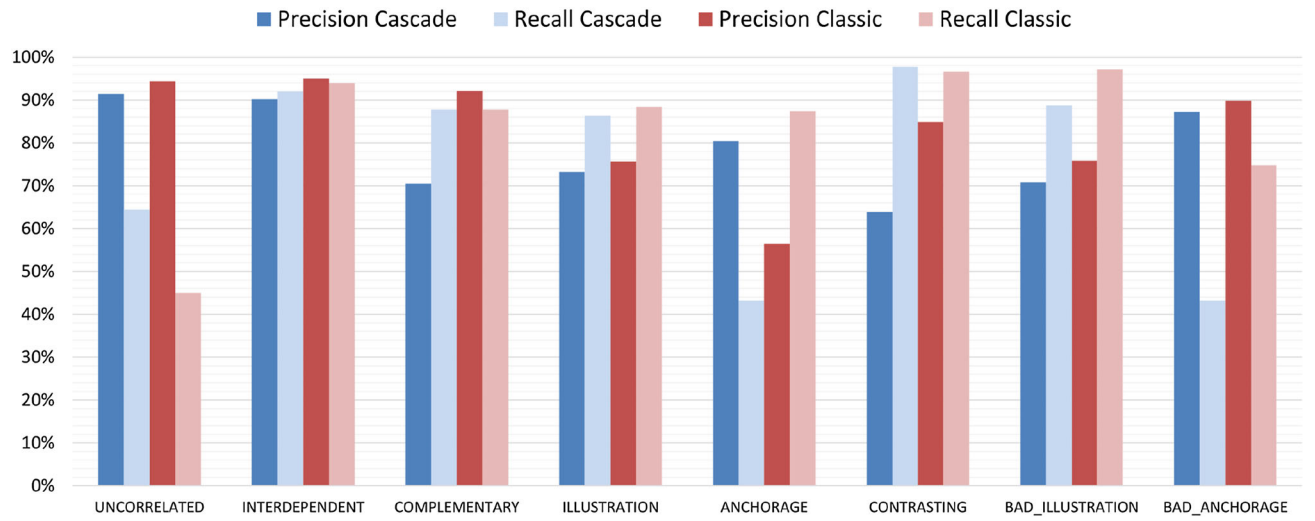

Fig. 9 Results for both classifiers

Table 6 Performance of the single metric classifiers

	CMI 0 (%)	CMI 1 (%)	–
Precision	87.72	91.40	–
Recall	80.32	94.90	–
	SC 0 (%)	SC 1 (%)	SC -1 (%)
Precision	81.79	84.21	86.63
Recall	90.51	64.43	88.38
	STAT 0 (%)	STAT T (%)	STAT I (%)
Precision	82.47	82.18	92.79
Recall	90.50	100.00	54.21

(Table 4) are presented in confusion matrices by means of precision and recall. For a better comparison, Fig. 9 shows the individual performance for each image-text class. The overall results for our classifiers in predicting CMI, SC, STAT as well as the image-text classes are presented in Table 7. The accuracy of the classifiers for CMI, SC and STAT ranges from 83.8 to 90.3%, while the two classification variations for the image-text classes achieved an accuracy of 74.3% (*cascade*) and 80.8% (*classic*). We also compared our method with our previous approach [13,14] by mapping their intermediate steps for CMI = 0, 1, 2 to 0, CMI=3,4 to 1, and SC = ±0.5 to ±1.

5.2 Discussion of results

As shown by Tables 4 and 5, the *classic* approach outperformed the *cascade* method by about 6% in terms of accuracy, indicating that a direct prediction of the image-text class is to be preferred over a combination of three separate classifiers. A reason might be that an overall judgment of the connection between image and text is probably more accurate than combining the independent ones, because all aspects of the multimodal message are regarded. This is also pleasant since an application would only need to train one classifier instead of three. Nonetheless, as can be seen in Table 7, results of the single metric classifiers suggest that they are still useful for applications that require just a single dimension, e.g., CMI for image captioning tasks. Regarding the image-text classes *Uncorrelated* achieved the lowest recall indicating that both classifiers often detected a connection (either in the SC dimension or CMI), even though there was none. This might be due to the concept detector contained in Inception-ResnetV2 focusing on negligible background elements that a human would not consider to be of importance (cf. Sect. 3.1). However, the high precision indicates that if it was detected it was almost always correct, in particular for the cascade classifier. The classes with positive SC are mainly confused with their negative counterparts, which is understandable since the difference between a positive and a negative SC is often caused by a few keywords in the text. But the performance is still impressive when considering that positive and negative

Table 7 Test set accuracy of the metric-specific classifiers and the two final classifiers after 75,000 iterations

Classifier	CMI (%)	SC (%)	STAT (%)	Cascade (%)	Classic (%)
Ours	90.3	84.6	83.8	**74.3**	**80.8**
[13]	68.8	49.6	–	–	–

instances differ only in a few keywords, while image content, sentence length and structure are identical.

The "cascaded" classifier struggled the most with both *Anchorage* classes, confusing them with *Complementary* and *Contrasting*. This is an indicator indicates that the Status classifier failed to identify that the text is subordinate and as can be seen in Table 6, it has indeed the lowest recall of the three dimensions.

Another interesting observation can be reported regarding the cascade approach: the rejection class *Undefined*, which is predicted if an invalid leaf of the categorization (the crosses in Fig. 4) is reached, can be used to judge the quality of our categorization. In total, 10 out of 18 leaves represent such an invalid case, but only 27 image-text pairs (3.4%) of all test samples were assigned to it. Thus, the distinction seems to be of high quality which is due to the good results of the classifiers for the individual metrics (Table 7).

Figure 10 shows some examples for correctly and incorrectly predicted image-text pairs. The third column in this Figure shows a false prediction of an uncorrelated pair as anchorage. There were some errors of false positives for anchorage (or illustrations), which seem to be partially caused by the typically corresponding shorter (or longer) text length. But the overall results indicate that the system does not solely rely on this feature, of course, otherwise a distinction of eight classes of this quality would not have been achievable. This is supported also by the correctly predicted example in Fig. 10, left, where despite the short text the image-text pair is classified as uncorrelated (and not as anchor).

6 Conclusions and future work

In this paper, we have presented a contribution to not only bridge the *semantic gap* between visual and textual information, but also the gap between research in linguistics and communication science on one side, and multimedia and computer vision research on the other side. By leveraging and extending the set of computable image-text metrics introduced in previous work [13], we have shown how they can be translated into intuitive, distinct semantic image-text classes. Our findings are motivated by research in linguistics and visual communication sciences, which identified similar classes. But instead of gathering distinct image-text classes through observation, which is common practice in those disciplines, we have derived image-text categories from our set of three metrics cross-modal mutual information, semantic correlation, and the status relation. We have further demonstrated how to (almost) automatically gather a dataset for the eight classes, which is then used to train baseline deep learning classifiers. We were able to predict the semantic image-text classes with an accuracy of 80.8%, while the accuracy was between 83% and 90% for the aforementioned metrics. We believe that the presented categorization and the automatic prediction of semantic image-text classes are a solid basis to enable a multitude of possible applications in fields such as multimodal Web content analysis and retrieval, cross-modal retrieval, or search as learning.

In the future, we plan to investigate additional metrics for image-text relations to further detail the identified classes. To do so, more Web resources need to be employed or potentially labeled manually. Finally, we will evaluate the usefulness of

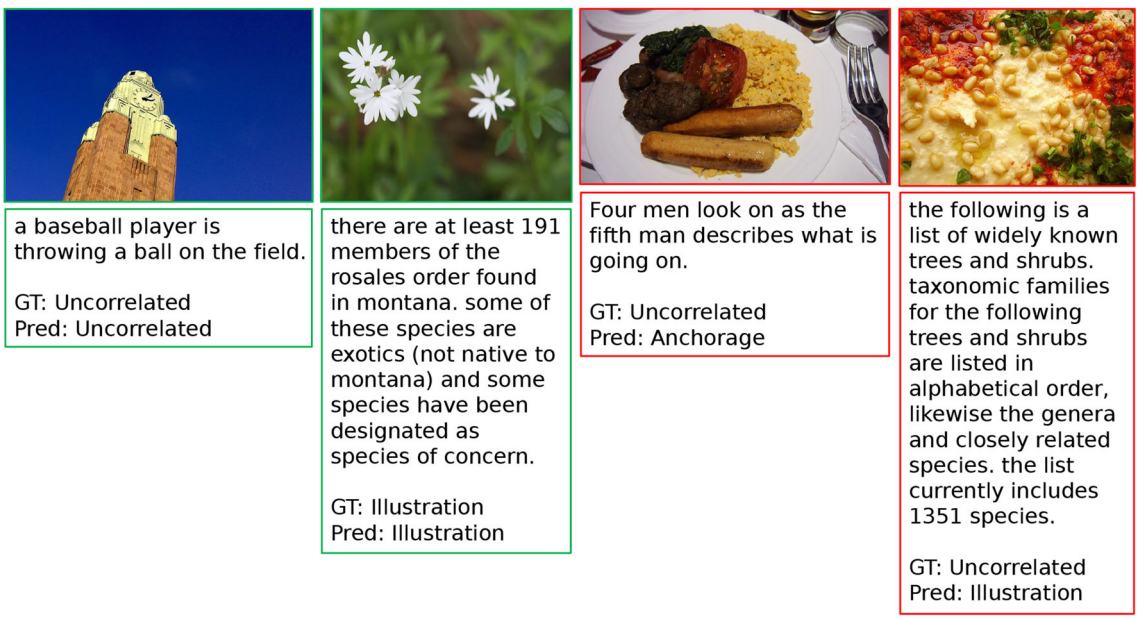

Fig. 10 Example predictions of the "classic" classifier. Green box: correct prediction; Red box: false prediction

our approach in different applications that can benefit from multimodal understanding, such as learning with multimedia data, retrieval applications, recommendations of advertisements, etc.

Acknowledgements Open Access funding provided by Projekt DEAL. Part of this work is financially supported by the Leibniz Association, Germany (Leibniz Competition 2018, funding line "Collaborative Excellence", Project SALIENT [K68/2017]).

References

1. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086
2. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Conference track proceedings 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015
3. Balaneshin-kordan S, Kotov A (2018) Deep neural architecture for multi-modal retrieval based on joint embedding space for text and images. In: Proceedings of the eleventh ACM international conference on web search and data mining. ACM, pp 28–36
4. Baltrusaitis T, Ahuja C, Morency L (2019) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443
5. Barthes R (1977) Image-music-text, vol 332. Fontana, London
6. Bateman J (2014) Text and image: a critical introduction to the visual/verbal divide. Routledge, London
7. Bucak SS, Jin R, Jain AK (2014) Multiple kernel learning for visual object recognition: a review. IEEE Trans Pattern Anal Mach Intell 36(7):1354–1369
8. Fan M, Wang W, Dong P, Han L, Wang R, Li G (2017) Cross-media retrieval by learning rich semantic embeddings of multimedia. In: Proceedings of the 2017 ACM conference on multimedia, MM 2017, Mountain View, CA, USA, October 23–27, 2017, pp 1698–1706
9. Gönen M, Alpaydin E (2011) Multiple kernel learning algorithms. J Mach Learn Res 12:2211–2268
10. Grave E, Mikolov T, Joulin A, Bojanowski P (2017) Bag of tricks for efficient text classification. In: Proceedings of the 15th conference of the european chapter of the association for computational linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017, Volume 2: Short Papers, pp 427–431
11. Halliday MAK, Matthiessen CM (2013) Halliday's introduction to functional grammar. Routledge, London
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp 770–778
13. Henning CA, Ewerth R (2017) Estimating the information gap between textual and visual representations. In: Proceedings of the 2017 ACM on international conference on multimedia retrieval, ICMR 2017, Bucharest, Romania, June 6–9, 2017, pp 14–22
14. Henning CA, Ewerth R (2018) Estimating the information gap between textual and visual representations. IJMIR 7(1):43–56
15. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, pp 2261–2269
16. Huang TK, Ferraro F, Mostafazadeh N, Misra I, Agrawal A, Devlin J, Girshick RB, He X, Kohli P, Batra D, Zitnick CL, Parikh D, Vanderwende L, Galley M, Mitchell M (2016) Visual storytelling. In: NAACL HLT 2016, The 2016 conference of the north american chapter of the association for computational linguistics: human language technologies, San Diego California, USA, June 12–17, 2016, pp 1233–1239
17. Hussain Z, Zhang M, Zhang X, Ye K, Thomas C, Agha Z, Ong N, Kovashka A (2017) Automatic understanding of image and video advertisements. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp 1100–1110
18. Jaques N, Taylor S, Sano A, Picard R (2015) Multi-task, multi-kernel learning for estimating individual wellbeing. In: Proceedings NIPS workshop on multimodal machine learning, Montreal, Quebec, vol 898
19. Johnson J, Karpathy A, Fei-Fei L (2016) Densecap: fully convolutional localization networks for dense captioning. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp 4565–4574
20. Karpathy A, Joulin A, Li F (2014) Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems 27: annual conference on neural information processing systems 2014, December 8–13 2014, Montreal, Quebec, Canada, pp 1889–1897
21. Kloepfer R (1977) Komplementarität von Sprache und Bild: am Beispiel von Comic, Karikatur und Reklame. Akad. Verlag-Gesell, Athenaion
22. Krippendorff K (1970) Estimating the reliability, systematic error and random error of interval data. Educ Psychol Meas 30(1):61–70
23. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012. proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., pp 1106–1114
24. Kruk J, Lubin J, Sikka K, Lin X, Jurafsky D, Divakaran A (2019) Integrating text and image: determining multimodal document intent in instagram posts. CoRR abs/1904.09073
25. Lan W, Li X, Dong J (2017) Fluency-guided cross-lingual image captioning. In: Proceedings of the 2017 ACM on multimedia conference, MM 2017, Mountain View, CA, USA, October 23–27, 2017, pp 1549–1557
26. Liang J, Li Z, Cao D, He R, Wang J (2016) Self-paced cross-modal subspace matching. In: Proceedings of the 39th international acm sigir conference on research and development in information retrieval, SIGIR 2016, Pisa, Italy, July 17–21, 2016, pp 569–578
27. Lin T, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. In: Proceedings computer vision-ECCV 2014–13th European conference, Zurich, Switzerland, September 6–12, 2014, Part V, pp 740–755

28. Liu F, Zhou L, Shen C, Yin J (2014) Multiple kernel learning in the primal for multimodal alzheimer's disease classification. IEEE J Biomed Health Inf 18(3):984–990
29. Marsh EE, Domas White M (2003) A taxonomy of relationships between images and text. J Doc 59(6):647–672
30. Martinec R, Salway A (2005) A system for image-text relations in new and old media. Vis Commun 4(3):337–371
31. Mazloom M, Rietveld R, Rudinac S, Worring M, van Dolen W (2016) Multimodal popularity prediction of brand-related social media posts. In: Proceedings of the 2016 ACM conference on multimedia MM 2016, Amsterdam, The Netherlands, October 15–19, 2016, pp 197–201
32. McCloud S (1993) Understanding comics: the invisible art. Northampton, Mass
33. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems 26: 27th annual conference on neural information processing systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, USA, pp 3111–3119
34. Mithun NC, Li J, Metze F, Roy-Chowdhury AK (2018) Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proceedings of the 2018 ACM international conference on multimedia retrieval. ACM, pp 19–27
35. Mithun NC, Panda R, Papalexakis EE, Roy-Chowdhury AK (2018) Webly supervised joint embedding for cross-modal image-text retrieval. In: Proceedings of the 26th ACM international conference on multimedia, MM '18. ACM, New York, NY, USA, pp 1856–1864
36. Nöth W (1995) Handbook of semiotics. Indiana University Press, Bloomington
37. Pages ME, List of antonyms and opposites (2017). http://www.myenglishpages.com/site_php_files/vocabulary-lesson-opposites.php. Accessed 23 Nov 2017
38. Poria S, Cambria E, Gelbukh AF (2015) Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, pp 2539–2544
39. Qi J, Peng Y, Zhuo Y (2018) Life-long cross-media correlation learning. In: 2018 ACM conference on multimedia, MM 2018, Seoul, Republic of Korea, October 22–26, 2018, pp 528–536
40. Ramanishka V, Das A, Park DH, Venugopalan S, Hendricks LA, Rohrbach M, Saenko K (2016) Multimodal video description. In: Proceedings of the 2016 ACM conference on multimedia, MM 2016, Amsterdam, The Netherlands, October 15–19, 2016, pp 1092–1096
41. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MS, Berg AC, Li F (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
42. Schifanella R, de Juan P, Tetreault JR, Cao L (2016) Detecting sarcasm in multimodal social platforms. In: Proceedings of the 2016 ACM conference on multimedia conference, MM 2016, Amsterdam, The Netherlands, October 15–19, 2016, pp 1136–1145
43. Shutova E, Kiela D, Maillard J (2016) Black holes and white rabbits: metaphor identification with visual features. In: NAACL HLT 2016, The 2016 conference of the north american chapter of the association for computational linguistics: human language technologies, San Diego California, USA, June 12–17, 2016, pp 160–170
44. Smeulders AWM, Worring M, Santini S, Gupta A, Jain RC (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
45. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, February 4–9, 2017, San Francisco, California, USA., pp 4278–4284
46. Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: IEEE Conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, pp 1–9
47. Unsworth L (2007) Image/text relations and intersemiosis: towards multimodal text description for multiliteracies education. In: Proceedings of the 33rd international systemic functional congress, pp 1165–1205
48. Van Leeuwen T (2005) Introducing social semiotics. Psychology Press, London
49. Xu N, Mao W (2017) Multisentinet: a deep semantic network for multimodal sentiment analysis. In: Proceedings of the 2017 ACM conference on information and knowledge management, CIKM 2017, Singapore, November 06–10, 2017, pp 2399–2402
50. Xu X, Song J, Lu H, Yang Y, Shen F, Huang Z (2018) Modal-adversarial semantic learning network for extendable cross-modal retrieval. In: Proceedings of the 2018 acm international conference on multimedia retrieval, ICMR 2018, Yokohama, Japan, June 11–14, 2018, pp 46–54
51. Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH (2016) Hierarchical attention networks for document classification. In: NAACL HLT 2016, The 2016 conference of the north american chapter of the association for computational linguistics: human language technologies, San Diego California, USA, June 12–17, 2016, pp 1480–1489
52. Yeh Y, Lin T, Chung Y, Wang YF (2012) A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection. IEEE Trans Multimed 14(3–1):563–574
53. Zhang M, Hwa R, Kovashka A (2018) Equal but not the same: Understanding the implicit relationship between persuasive images and text. In: British machine vision conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3–6, 2018, p 8