Leveraging High Level Visual Information for Matching Images and Captions

Fei Yan and Krystian Mikolajczyk

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, Surrey, GU2 7XH, UK
{f.yan, k.mikolajczyk}@surrey.ac.uk

Abstract. In this paper we investigate the problem of matching images and captions. We exploit the kernel canonical correlation analysis (KCCA) to learn a similarity between images and texts. We then propose methods to build improved visual and text kernels. The visual kernels are based on visual classifiers that use responses of a deep convolutional neural network as features, and the text kernel improves the Bag-of-Words (BoW) representation by learning a vision based lexical similarity between words. We consider two application scenarios, one where only an external image set weakly related to the evaluation dataset is available for training the visual classifiers, and one where visual data closely related to the evaluation set can be used. We evaluate our visual and text kernels on a large and publicly available benchmark, where we show that our proposed methods substantially improve upon the state-of-the-art.

1 Introduction

The explosive growth of visual and textual data on the web and in personal collections demands effective methods for image and video search, visual content description and text-to-image generation. Owing to recent advances in computer vision (CV), natural language processing (NLP) and machine learning (ML), integrated modelling of vision and language is finding more and more applications, e.g. face recognition from caption-based supervision [1], text-to-image coreference [2], and zero-shot visual learning using purely textural description [3]. In particular, generating natural language description for image and video has attracted much interest in both CV and NLP communities [4–16].

One of the main issues with the work in [4–16], however, is the lack of automatic and objective evaluation metric. In [17] the problem of generating natural language description for a given image is relaxed to one of ranking a set of human-written captions, by assuming the set contains the original (human-written) caption of the image. [17] builds a dataset (dubbed Flickr8K) of image and caption pairs, and employs the kernel canonical correlation analysis (KCCA) [18, 19] to learn a latent space in which a similarity measure between an image and a caption is defined. KCCA requires two kernels to be built, one for the images and the other for the captions. [17] fixes the image kernel to a relatively simple one

that uses only low level and mid-level visual information such as colour, texture and SIFT descriptors, and demonstrates that text kernels that exploit lexical similarities and high-order co-occurrence information outperform the basic Bag-of-Word (BoW) text kernel.

In this paper, we build on the results from [17], and propose an approach that significantly improves the performance of image-to-text annotation and text-to-image retrieval. In particular we consider the scenario where there is no image data for training visual classifiers for synsets in the application domain at hand. Our contributions can be summaries as follows:

- Our approach makes use of additional visual classifiers for synsets different than those contained in the evaluation data. We show that the combination of the basic BoW text kernel and a high level image kernel based on the probabilities given by visual classifiers outperforms the best combination in [17] in most evaluation metrics (Section 5.1);
- We demonstrate that visual classifiers trained for synsets included in the evaluation dataset improve the retrieval scores by a factor of two compared to the best method in [17] (Section 5.2);
- Finally, in contrast to lexical similarities computed using text corpora, we propose to use the high level visual information to learn a lexical similarity, and show that the BoW text kernel enriched with such lexical similarity further boosts the performance (Section 6).

The remainder of this paper is organised as follows: in Section 2 we brief review existing work on description generation for image and video. In particular, we present the dataset and experimental setup introduced in [17], which we compare our approaches to. The BoW text kernel used in our studies is presented in Section 3, followed by an introduction in Section 4 to our visual recognition pipeline that is based on a deep convolutional neural network (CNN). In Section 5, the performance of high level visual kernels built using external and internal sets of synsets is presented respectively, with analysis and discussions. In Section 6, we propose a vision based lexical similarity to model partial matches in the text representation, and compare it to language based lexical similarities. Finally, Section 7 concludes the paper.

2 Related Work

Generating natural language description for image and video has become a popular research topic in recent years. Among existing work on this topic, the goal in [4] is automatic caption generation for a given news image with an associated news article. A good caption for such an image is often only loosely related to the content of the image. The setting of this work is therefore different from that in [5–12], where the objective is to generate a caption that describes what is depicted in the image.

In [5], a dataset with 1 million image-caption pairs is leveraged, and the caption of the image in the dataset that is visually most similar to the given image

is transferred as its caption. However, even with 1 million images, it is unrealistic to expect that every possible query image with various objects and actions can be represented and found in such dataset. In contrast to this caption transfer approach, the work in [6–12] adopts the conventional content selection and surface realisation approach. Starting from the output of visual processing engines e.g. object classifiers, object detectors and attribute classifiers, image content that will be described is selected in the form of tuples such as subject-action-object triplets, object-preposition-object triplets, and object-action-preposition-scene triplets quadruplet. A surface realiser is then employed to produce captions as constrained by the lexicon and grammar.

While [6] focuses on the investigation of surface realisation techniques, the work in [7–12] differs primarily in the way the tuples of image content are generated. In [7, 8], structured output learning techniques i.e. structured support vector machine (S-SVM) and conditional random field (CRF) are employed to learn the mapping from the output of visual processing engines to the tuples. In [11] the tuples for the test image are a weighted collection of those for the training examples, and the weights are learnt from training image-caption pairs. In [12] content selection and planning is formulated as an integer linear program (ILP). Finally, [9, 10] employ language corpora to learn word co-occurrence statistics and use the statistics to filter and enrich the output of visual processing engines. While [9] uses a hidden Markov model (HMM) to determine the quadruplet of image content, [10] generates syntactic trees to encode what the visual engines see.

In parallel to image captioning, automatic video description is also receiving increasing attention [13–16]. Although details differ, [13–16] operate within the same paradigm of content selection and surface realisation. Compared to image description, typically video description systems additionally employ spatio-temporal methods for action recognition.

2.1 Image Description as a Ranking Task

The approaches discussed above can produce human-like description for image and video. However, [17] argues that these approaches lack automatic and objective evaluation methods. On the one hand, although automatic evaluation metrics such as BLEU [20] and ROUGE [21] are useful for measuring the fluency of the generated text [22], it is shown in [17] that they are not reliable metrics for how accurately a caption describes an image. On the other hand, human judgements can be quite reliable but are expensive and time-consuming to collect. To address this issue, [17] builds a dataset of image-caption pairs, and formulates the image captioning problem as one of ranking a set of available human-written captions.

The Flickr8K dataset The authors of [17] collected 8000 images from the Flickr.com website, which focused on people or animal performing actions. Using a crowdsourcing service, five captions were generated by different annotators for each image. The annotators were asked to describe the actors, objects, scenes and

	– A cat standing on carpet is interested in a piece of string in the air nearby wood flooring. – A white and brown cat bats at a frayed string dangling in front of him. – Cat playing with a dangling string. – Cat standing to play with string. – The white and black cat pawed at the piece of fabric.
	– A couple of several people sitting on a ledge overlooking the beach. – A group of people sit on a wall at the beach. – A group of teens sit on a wall by a beach. – Crowd of people at the beach. – Several young people sitting on a rail above a crowded beach.

Table 1. Two example image-caption pairs in the Flickr8K dataset.

activities that were shown in the image, i.e., information that could be obtained from the image alone. Two examples of image-caption pairs are illustrated in Table 1. The dataset is split into predefined training, validation, and test sets with 6000, 1000, and 1000 pairs respectively. The five captions are pooled into one for the training set, and in the validation and test sets only caption 2 is used. Each image-caption pair therefore can be thought of as consisting of one image and one caption.

Kernel canonical correlation analysis (KCCA) Given m samples of two sets of variables $\{\mathbf{x}_i\}_{i=1}^m$ and $\{\mathbf{y}_i\}_{i=1}^m$ where $\mathbf{x}_i \in \mathcal{R}^{d_x}$ and $\mathbf{y}_i \in \mathcal{R}^{d_y}$. Canonical correlation analysis (CCA) [23] finds a projection for each set such that in the projected common space the linear correlation of the samples is maximised. CCA can be kernelised by implicitly embedding $\mathbf{x}_i$ and $\mathbf{y}_i$ into feature spaces through kernel functions $k_x(\mathbf{x}_i, \mathbf{x}_j)$ and $k_y(\mathbf{y}_i, \mathbf{y}_j)$ [18, 19]. The resulting kernel CCA (KCCA) finds the two projections by solving:

$$\underset{\alpha,\beta}{\text{argmax}} \frac{\alpha^T K_x K_y \beta}{\sqrt{(\alpha^T K_x^2 \alpha + \kappa \alpha^T K_x \alpha)(\beta^T K_y^2 \beta + \kappa \beta^T K_y \beta)}} \tag{1}$$

where K_x and K_y are the $m \times m$ training kernel matrices with $K_x[i,j] = k_x(\mathbf{x}_i, \mathbf{x}_j)$ and $K_y[i,j] = k_y(\mathbf{y}_i, \mathbf{y}_j)$, and κ is a regularisation parameter. Let l be the number of test examples, and K'_x and K'_y be the $m \times l$ test kernel matrices. The similarity measure between test image i' and test caption j' is defined as the cosine of the angle between the two projected points in the learnt common space:

$$\text{Sim}(\mathbf{x}_{i'}, \mathbf{y}_{j'}) = \cos(\alpha^T K'_x[:, i'], \beta^T K'_y[:, j']) \tag{2}$$

or the linear correlation between them:

$$\text{Sim}(\mathbf{x}_{i'}, \mathbf{y}_{j'}) = \text{corr}(\alpha^T K'_x[:, i'], \beta^T K'_y[:, j']) \tag{3}$$

where $K'_x[:, i']$ denotes the i'^{th} column of K'_x, and $K'_y[:, j']$ denotes the j'^{th} column of K'_y.

Evaluation metrics Given an image kernel and a text kernel, [17] learns the common space using KCCA. For each test image, the 1000 captions in the test set are ranked according to their similarity to the image using Eq. (2). This ranked list allows to define metrics that measure how well images and captions are matched in the learnt common space. [17] proposes to use the recall of the caption originally paired with the image at position 1, 5, 10 of the ranked list (R@1, R@5, R@10), and the median rank (MR) of this original gold caption for all test images.

Formulating caption generation as a ranking task allows different approaches to be compared in an automatic, efficient and objective fashion. Moreover, such a framework can be trivially extended to perform the symmetric task of image retrieval using captions. We therefore employ the KCCA approach for both image-to-text annotation and text-to-image retrieval. We use the metrics R@1, R@5, R@10 and MR to measure the performance of our kernels on both annotation and retrieval tasks, and compare with that in [17].

3 Bag-of-Words (BoW) Text Kernel

In this section, we introduce the text kernel K_y used in our study. First, all captions are processed using the linguistic analyser of [24]. This analyser performs tokenisation, lemmatisation, part-of-speech (POS) tagging and word-sense disambiguation. There are 5768 unique lemmatised words in the training captions. We build a $d_y = 5768$ dimensional bag-of-words (BoW) representation for each caption, with the r^{th} dimension:

$$y^r = t^r \log \frac{D}{d^r + 1} \tag{4}$$

where t^r is the term frequency (TF) of the r^{th} lemmatised word i.e. the number it appears in the caption, d^r is document frequency of the lemmatised word i.e. the number of training captions where it appears, and D is the total number of training captions. We adopt the linear correlation as the kernel function:

$$k_y(\mathbf{y}_i, \mathbf{y}_j) = \text{corr}(\mathbf{y}_i, \mathbf{y}_j) \tag{5}$$

and denote the resulting kernel $BoW5'$, where "5" indicates that for the training set the five captions are pooled into one, and the prime symbol indicates that $BoW5'$ is a close variant of the $BoW5$ kernel used in [17] with the following differences: 1) $BoW5'$ adopts the linear correlation as kernel function while $BoW5$ adopts the cosine; 2) $BoW5'$ uses the standard inverse document frequency (IDF) weight while $BoW5$ uses a square-rooted version which is found to perform better in [17]; 3) stop words are kept when building $BoW5'$ while they are removed in $BoW5$.

Table 2. Statistics of the sets of synsets involved in our work. Visual classifiers are trained for the two sets in boldface.

Synset set	# of synsets	# of images
{ILSVRC12}	1000	1,281,169
{Caption}	3335	-
{ImageNet}	21841	-
{ILSVRC12} ∩ {ImageNet}	999	-
{ILSVRC12} ∩ {Caption}	197	-
{Caption} ∩ {ImageNet}	1372	1,571,576

The word-sense disambiguation component of the linguistic analyser also maps each token to a WordNet synset. We denote by {Caption} the set of 3335 synsets that correspond to tokens in the captions labelled as nouns by the POS tagger. In the following, we train visual classifiers for a subset of {Caption}, and use the output of the classifiers to build a high level visual kernel.

4 Building High Level Visual Kernels with Deep Learning

Following the success of deep convolutional neural network (CNN) [25] in the ImageNet large scale visual recognition challenge 2012 (ILSVRC12) [26], deep learning [27, 28] has become the de-facto approach for large scale visual recognition. Moreover, it has recently been shown in [29] that features extracted from the activations of a deep CNN trained in a fully supervised fashion can be repurposed to novel generic tasks that differ significantly from the original task.

Inspired by [29], we extract such activations as features for novel visual recognition tasks. More specifically, we train binary classifiers for two sets of WordNet synsets: the first set, denoted {ILSVRC12}, is the synsets in the ILSVRC12 challenge; and the second set, denoted {Caption} ∩ {ImageNet}, is the intersection of the 3335 noun synsets in the captions of Flickr8K and the 21841 synsets in the 2011 Fall release of the ImageNet [30]. Statistics of the sets involved are summarised in Table 2.

For each of the two sets, we extract activations of a pre-trained CNN model as features for images in the synsets. Similarly, features are also extracted for the images in Flickr8K. The pre-trained CNN model is a reference implementation of the structure proposed in [25] with minor modifications, and is made publicly available through the Caffe project [31]. It is shown in [29] that the activations of layer six of the CNN perform the best for novel tasks. Our study on a toy example with ten ImageNet synsets however suggests that the activations of layer seven have a small edge.

Again for each of the two sets, once the 4096 dimensional activations of layer seven are extracted, we train binary support vector machines (SVMs) using the LIBSVM toolbox [32] for each synset, with 5000 images randomly sampled from

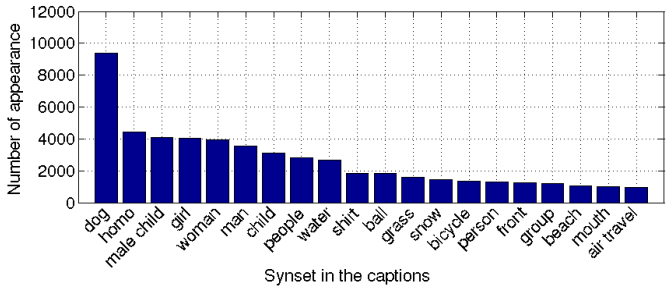


Fig. 1. The top 20 most frequently appearing synsets in Flickr8K. X-axis: the first word/phrase in the WordNet definition of a synset; Y-axis: number of appearances in the $8000 \times 5 = 40000$ original captions (before pooling for training set).

all synsets as negative examples. The trained SVMs are used to predict the probabilities of the presence of the n synsets in the Flickr8K images.

Let $\mathbf{x}$ be the $d_x = n$ dimensional representation of an image, where its t^{th} element x^t is the probability that synset t is present in the image, as given by the t^{th} SVM. We again use the linear correlation as kernel function:

$$k_x(\mathbf{x}_i, \mathbf{x}_j) = \text{corr}(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

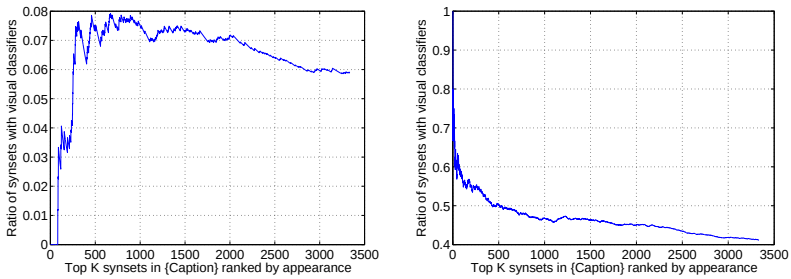
Compared to the visual kernel in [17] which uses only low and mid- level visual information such as colour, texture and SIFT descriptors, our kernel encodes high level visual information in terms of presence of objects, actions, and scenes.

5 Evaluation of High Level Visual Kernels

In this section, we evaluate the high level visual kernels in conjunction with the $BoW5'$ text kernel under the KCCA framework, and provide analysis and discussions on the results. To enable a fair comparison, we follow [17] and find 15 best performing models on the validation set by tuning the KCCA regularisation parameter κ and the dimensionality d of the learnt common space. The final rank on the test set is obtained by aggregating the ranks given by the 15 sets of optimal parameters. In the following, we consider two scenarios: when visual classifiers are leant for sysnets that are external to the Flickr8K dataset i.e. synsets in {ILSVRC12} (Section 5.1); and when they are learnt for synsets from the captions of Flickr8K i.e. synsets in {Caption} $\cap$ {ImageNet} (Section 5.2).

5.1 Learning for Synsets in {ILSVRC12}

To build the Flickr8K dataset, images were collected from six Flickr groups: *strangers!*, *Wild-Child*, *Dogs in Action*, *Outdoor Activities*, *Action Photography*, and *Flickr-Social*. As a result, the images tend to depict people or animals (mainly dogs) performing some action. The top 20 most frequently appearing



(a) {ILSVRC12} (b) {Caption} ∩ {ImageNet}

Fig. 2. Ratio of top K synsets in {Caption} ranked by appearances that have visual classifiers. (a): when learning for synsets in {ILSVRC12}; (b): when learning for synsets in {Caption} ∩ {ImageNet}.

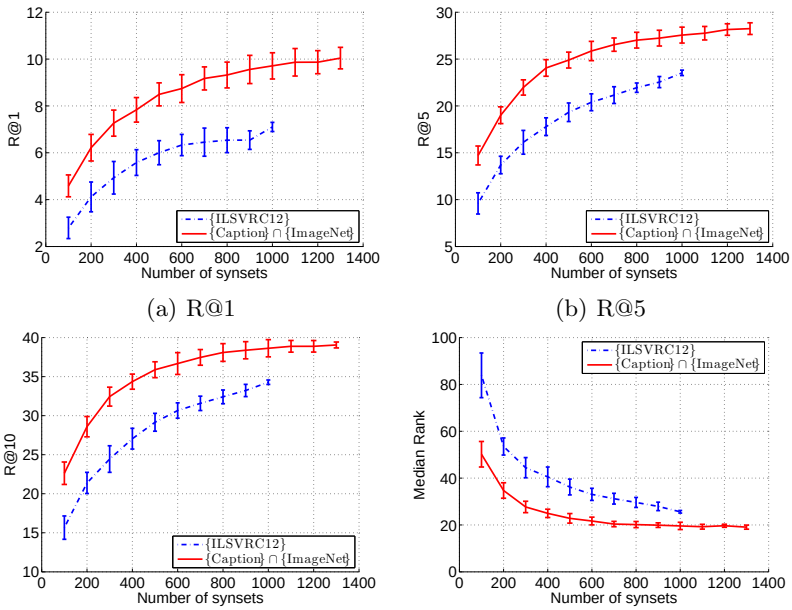
Table 3. Performance of high level visual kernel learnt on {ILSVRC12}, ILS.

		Image annotation				Image retrieval			
Image	Text	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
Pyr	$BoW5$	6.2	17.1	24.3	58.0	5.8	16.7	23.6	60.0
Pyr	$TagRank$	6.0	17.0	23.8	56.0	5.4	17.4	24.3	52.5
Pyr	$Tri5$	7.1	17.2	23.7	53.0	6.0	17.8	26.2	55.0
Pyr	$Tri5Sem$	**8.3**	21.6	30.3	34.0	**7.6**	20.7	30.1	38.0
ILS	$BoW5'$	7.5	**21.8**	**33.3**	**26.0**	6.6	**24.6**	**34.7**	**26.0**

synsets in the set {Caption} are shown in Fig. 1, where we can see that synset *dog, domestic dog, Canis familiaris* (WordNet ID n02084071) appears in more than 9000 captions out of the original 40000, twice more than synset *homo, man, human being, human* (WordNet ID n02472293). Note that to avoid clutter, in Fig. 1 as well as in the rest of this paper, we use the first word of the WordNet definition to refer to a synset, e.g., we use *dog* instead of *dog, domestic dog, Canis familiaris* for synset n02084071.

{ILSVRC12} provides training images for synsets external to Flickr8K and has a very small intersection with {Caption}. According to Table 2, the two sets have 197 common elements. More details are presented Fig. 2 (a), which shows the ratio of the top K synsets in {Caption} that have visual classifiers (i.e. also in {ILSVRC12}). The first synset with a visual classifier appears at $K = 83$, and the fraction of Flickr8K noun synsets with visual classifiers is lower than 8%. The low level of overlapping between {ILSVRC12} and {Caption} fits our application scenario where little directly related training data is available.

In Table 3 we present the performance of our high level visual kernel ILS that is built by visual classifiers for the synsets in {ILSVRC12}. For comparison we also include the performance of the methods reported in [17], where Pyr denotes the visual kernel in [17] that uses only low and mid- level visual information, $TagRank$, $Tri5$ and $Tri5Sem$ are sophisticated text kernels that use high-order



(a) R@1 (b) R@5

(c) R@10 (d) Median Rank

Fig. 3. Mean and standard deviation of performance averaged over "annotation" and "retrieval" tasks as functions of the number of random synsets.

word statistics and lexical similarities. Table 3 shows the recalls R@1, R@5, R@10 and the median rank on both annotation and retrieval tasks.

The results in Table 3 demonstrate that ILS performs well despite the small overlap between {ILSVRC12} and {Caption}. When the basic text kernels $BoW5$ or $BoW5'$ are used, ILS outperforms Pyr by large margins in all metrics and on both tasks. For example, with the same Pyr visual kernel, the text kernels $TagRank$ and $Tri5$ that exploit high-order word statistics reduce the median rank of $BoW5$ from 58 to 56 and 53 respectively, while ILS takes it to 26. This suggests that compared to noisy low level visual representation, better alignment can be found between high level visual information and captions.

When compared to the best combination in [17] $Pyr/Tri5Sem$, where $Tri5Sem$ encodes both third-order word co-occurrence statistics and lexical similarities learnt from several external corpora, the $ILS/BoW5'$ combination still leads in most metrics. Interestingly, while $Tri5Sem$ seems better at finding the exact gold item as indicted by the higher R@1 scores, ILS can bring good matches to the query overall, leading to better R@5, R@10 and median rank scores.

We also randomly sample 100 to 900 synsets at a step size of 100 and build visual kernels using only the predicted probabilities from the corresponding SVMs. For each sample size we repeat the experiment 15 times, and report the mean and standard deviation of the four metrics in Fig. 3. Note that in Fig. 3 the performance has been averaged over the two tasks.

Table 4. Performance of high level visual kernel learnt on {Caption} ∩ {ImageNet}, $CapIma$.

		Image annotation				Image retrieval			
Image	Text	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
Pyr	$Tri5Sem$	8.3	21.6	30.3	34.0	7.6	20.7	30.1	38.0
ILS	$BoW5'$	7.5	21.8	33.3	26.0	6.6	24.6	34.7	26.0
$CapIma$	$BoW5'$	**10.0**	**27.3**	**39.2**	**19.0**	**9.8**	**28.8**	**38.6**	**19.0**

It is clear that the performance curves in Fig. 3 show no sign of saturation, which suggests that with visual classifiers trained for more sysnets the performance can be further improved. These results confirm that learning high level visual representation for a set of sysnets external to the captions is a viable approach for matching images and captions. This can be very useful as in practice training images are not always available for all synsets in the captions.

5.2 Learning for Synsets in {Caption} ∩ {ImageNet}

In the second scenario, we learn for synsets that are actually in the captions. To obtain training images we use the intersection of {Caption} and {ImageNet}. The resulting set {Caption} ∩ {ImageNet} has 1372 synsets with image data available for training visual classifiers.

Fig. 2 (b) plots the ratio of the top K synsets in {Caption} that have visual classifiers. Compared to Fig. 2 (a) the ratio here is much higher. For example, out of 10, 50, and 100 top ranked synsets, 7, 31, and 57 respectively have corresponding visual classifiers. Overall, 1372 synsets out of the total 3335 have visual classifiers.

In Table 4 we report the performance of the visual kernel $CapIma$ that is built using {Caption}∩{ImageNet}. For convenience we also repeat the results of $Pyr/BoW5$ and $ILS/BoW5'$ from Table 3. Table 4 shows that $CapIma/BoW5'$ combination outperforms the other two by large margins in all metrics and on both tasks. For example, its median rank (19.0) almost halves that of $Pyr/BoW5$ (36.0), which is the best reported in [17]. The significant edge of $CapIma/BoW5'$ over $ILS/BoW5'$ demonstrates the advantage of learning the visual appearance of the objects, scenes and actions that are actually mentioned in the captions, rather than learn the context.

Fig. 3 plots the performance when using varying numbers of randomly sampled synsets to build visual kernels, where for each number the averaged result for 15 random sets is reported. It gives a flavour of the performance that can be expected when training images are available only for limited number of synsets. For instance, when only 250 synsets have visual classifiers, the R@10 score is approximately 30, which is similar to that of $Pyr/Tri5Sem$.

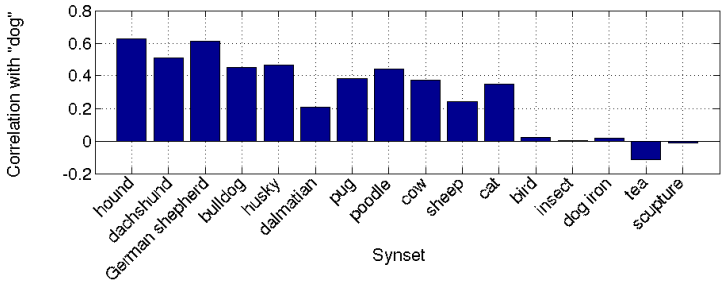


Fig. 4. Visual correlation with *dog* synset.

6 Vision Based Lexical Similarity

One of the key problems with the basic $BoW5/BoW5'$ text kernels is that they require exact match of words, and cannot account for the fact that the same entity can be described in different words. In [17] three lexical similarity measures are leant on text corpora. These similarities capture semantic relatedness, hence allow partial matches between words. The $Tri5Sem$ text kernel is then built by combining the lexical similarities, and is shown to be the best performing text kernel.

In contrast to the linguistics based similarities, we propose a vision based lexical similarity measure. This measure exploits the high level visual information encoded in the output of the visual classifiers. Recall that x^t in the representation of an image is the prediction of the visual classifier corresponding to the t^{th} synset. Let $\mathbf{x}^t \in \mathcal{R}^m$ and $\mathbf{x}^{t'} \in \mathcal{R}^m$ be predictions of the presence of synsets t and t' for all training images in Flickr8K. The linear correlation $c(t, t') = \text{corr}(\mathbf{x}^t, \mathbf{x}^{t'}) \in [-1, 1]$ can be thought of as a visually informed lexical similarity between synsets t and t'.

Fig. 4 plots the correlations between the *dog* synset and another 16 synsets. The figure shows that all breeds of dog have high positive correlations with *dog*. The correlations between the three mammals *cow*, *sheep*, *cat* and *dog* are also high, although in general not as high as the dog breeds. On the other hand, *bird*, *insect*, and the semantically unrelated ones all correlate poorly with *dog*. This demonstrates the potential advantages of the vision based lexical similarity.

The Lin similarity [33] used in [17] exploits the hypernym/hyponym relations in WordNet. As a result, synsets that have close relations but are not visually similar may have high similarity, for example, *dog* and *bird*, *swimming* and *football*. This particularly poses a problem when alignment between image and text is sought. Our vision based similarity measure, on the the hand, tackles the very problem. Moreover, the vision based similarity is not confused by the presence of words in semantically unrelated sysnets, e.g. *dog* in *dog iron*, which are visually dissimilar.

Recall that the word-sense disambiguation component of the linguistic analyser of [24] establishes correspondences between the lemmatised words and synsets. We consider the case where visual classifiers are trained for synsets in $\{\text{Caption}\}\cap$

Table 5. Performance of vision based lexical similarity.

		Image annotation				Image retrieval			
Image	Text	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
Pyr	$Tri5Sem$	8.3	21.6	30.3	34.0	7.6	20.7	30.1	38.0
ILS	$BoW5'$	7.5	21.8	33.3	26.0	6.6	24.6	34.7	26.0
$CapIma$	$BoW5'$	10.0	27.3	39.2	19.0	9.8	28.8	38.6	19.0
$CapIma$	$BoW5'_V$	**11.1**	**29.8**	**42.2**	**16.0**	**11.2**	**30.7**	**40.9**	**15.0**

{ImageNet}. Let syn(r) be the synset ID in {Caption}$\cap${ImageNet} that corresponds to the r^{th} word in the dictionary of 5768 unique words. The r^{th} dimension of the BoW representation with vision based similarity incorporated, is then:

$$y^r = \left(t^r + \gamma \sum_{s \in \{S\} \setminus r} t^s c\big(\text{syn}(r), \text{syn}(s)\big) \right) \log \frac{D}{d^r + 1} \tag{7}$$

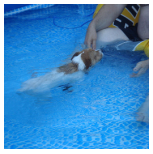
where $\{S\}$ is the set of word IDs in the dictionary whose corresponding sysnets have visual classifiers, t^s is the term frequency of the s^{th} word in the dictionary, and $\gamma \in [0, 1]$ is a parameter that is learnt on the validation set. When γ is set to 0, Eq. (7) reduces to the standard BoW representation in Eq. (4). Otherwise, the additional term accounts for partial matches between words: not only the r^{th} word activates the r^{th} bin of the BoW representation, other words also contribute an amount determined by the visual similarity between them and the r^{th} word.

We denote by $BoW5'_V$ the text kernel using vision based lexical similarity and report its performance in Table 5, where for comparison we repeat the content of Table 4. The $CapIma/BoW5'_V$ combination outperforms the previous best $CapIma/BoW5'$ in all metrics, reducing the median rank from 19.0 to 15.5. This clearly indicates the advantage of exploiting vision based lexical similarity.

The $Tri5Sem$ kernel in [17] combines several linguistics based lexical similarities trained on both internal (captions) and external corpora. Among them, the alignment based similarity takes advantage of the fact that each image in Flickr8K is associated with five independently written captions. Our vision based similarity does not rely on this and is therefore more general. Moreover, note that ideally we would like to compute the visual similarity between every pair of noun synsets, or even ever pair of sysnets, in the captions. In practice, however, we are constrained by available training images, and as a result we have pairwise similarity only for 1372 synsets. The potentials of the vision based lexical similarity are therefore not fully realised.

Finally, the five top ranked and the gold captions/images for three random test examples on both tasks are shown in Table 6 and Table 7, where the best kernels $CapIma/BoW5'_V$ are used.

- A boy in a park playing with two orange balls.
- little girls in swimsuits are laughing
- A boy in a bathing suit stands in water.
- A dark man in a white and green feathered mask with green jewelry and pants.
- A child wearing swim goggles.
- . . .
- **A man and a woman in festive costumes dancing.**



- A black dog in water.
- A child slides down a slide and into the water.
- A boy is diving through the air into a swimming pool.
- a person doing the backstroke in a swimming pool
- **a brown and white dog swimming towards some in the pool**

- A basketball player wearing a black and white uniform dribbles the ball.
- A boy with red shorts is holding a basketball in a basketball court.
- A basketball player dribbles the ball while another blocks him and an official looks on.
- Several basketball players are grabbing for the ball during a game.
- A basketball game
- . . .
- **A player from the white and green highschool team dribbles down court defended by a player from the other team.**

Table 6. Query image, the five top ranked captions retrieved (from top to bottom), and the gold caption (in boldface). In the three random examples the rank of the gold caption is 9, 5, and 11 respectively.

The dogs are in the snow in front of a fence.

A hiker ascends a snowy hill.

Three boys in a building under construction.

Table 7. Query caption, the five top ranked images retrieved (from left to right), and the gold image (in column 6). In the three random examples the rank of the gold image is 22, 16, and 810 respectively.

7 Conclusions

We have presented an approach for matching images and captions based on KCCA. Our visual kernels encode high level visual information resulting from state-of-the-art image recognition, leading to a significant improvement compared to low level visual representation in [17]. We successfully make use of additional annotated data with very few labels directly related to the test images, and we quantify the gain in performance when the visual classifiers are trained for directly related synsets. We have also proposed to exploit responses of visual classifiers to compute a lexical similarity between words. We evaluated the proposed approaches on a large and publicly available dataset, and showed that our methods substantially improved the state-of-the-art performance.

Acknowledgement

This work has been supported by EU Chist-Era EPSRC EP/K01904X/1 Visual Sense project.

References

1. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Face recognition from caption-based supervision. IJCV **96(1)** (2012) 64–82
2. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? text-to-image coreference. In: CVPR. (2014)
3. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textural description. In: ICCV. (2013)
4. Feng, Y., Lapata, M.: Automatic caption generation for news images. PAMI **35(4)** (2013) 797–812
5. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: NIPS. (2011)
6. Li, S., Kulkarni, G., Berg, T., Berg, A., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: CoNLL. (2011)
7. Farhadi, A., Hejrati, M., Sadeghi, M., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences for imgages. In: ECCV. (2010)
8. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A., Berg, T.: Baby talk: Understanding and generating simple image descriptions. In: CVPR. (2011)
9. Yang, Y., Teo, C., III, H.D., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: EMNLP. (2011)
10. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Daume, H.: Midge: Generating image descriptions from computer vision detections. In: EACL. (2012)
11. Gupta, A., Verma, Y., Jawahar, C.: Choosing linguistics over vision to describe images. In: AAAI Conference on Artificial Intelligence. (2012)
12. Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., Choi, Y.: Collective generation of natural image descriptions. In: ACL. (2012)

13. Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., Guadarrama, S.: Generating natural-language video descriptions using text-mined knowledge. In: AAAI Conference on Artificial Intelligence. (2013)
14. Das, P., Xu, C., Doell, R., Corso, J.: A thousand frames in just a few words: Lingual description of videos through latent topic and sparse object stitching. In: CVPR. (2013)
15. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: ICCV. (2013)
16. Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B.: Translating video content to natural language descriptions. In: ICCV. (2013)
17. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47** (2013) 853–899
18. Bach, F., Jordan, M.: Kernel independent component analysis. JMLR **3** (2002) 1–48
19. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation **16(12)** (2004) 2639–2664
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: ACL. (2002)
21. Lin, C.: ROUGE: a package for automatic evaluation of summaries. In: Workshop on Text Summarization Branches Out. (2004)
22. Reiter, E., Belz, A.: An investigation into the validity of some metrics for automatically evaluating natural lanugage generation systems. Computational Linguistics **35(4)** (2009) 529–338
23. Hotelling, H.: Relations between two sets of variates. Biometrika **28(3/4)** (1936) 321–377
24. Padro, L., Stanivlosky, E.: Freeling 3.0: Towards wider multilinguality. In: Language Resources and Evaluation Conference. (2012)
25. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: NIPS. (2012)
26. Deng, J., Berg, A., Satheesh, S., Su, H., Khosla, A., Feifei, L.: ImageNet large scale visual recognition challenge (ILSVRC) 2012. http://image-net.org/challenges/LSVRC/2012/ (2012)
27. LeCun, Y., Boser, B., denker, J., Henerson, D., Howard, R., Hubbard, W., Jackel, L.: Backpropagation applied to handwritten zip code recognition. Neural Computation **1(4)** (1989) 541–551
28. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313** (2006) 504–507
29. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. arXiv:1310.1531 [cs.CV] (2013)
30. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large scale hierarchical image database. In: CVPR. (2009)
31. Jia, Y.: Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org (2013)
32. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2(3)** (2011) 1–27 Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
33. Lin, D.: An information-theoretic definition on similarity. In: ICML. (1998)