Computer Vision and Image Understanding 000 (2017) 1–11

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network

T. Bui[a,*], L. Ribeiro[b], M. Ponti[b], J. Collomosse[a]

[a] Centre for Vision, Speech and Signal Processing (CVSSP) University of Surrey, Guildford, GU2 7XH, United Kingdom
[b] Institute of Mathematical and Computer Sciences (ICMC) Universidade de São Paulo, São Carlos/SP, 13566-590, Brazil

ARTICLE INFO

Article history:
Received 8 June 2016
Revised 1 April 2017
Accepted 19 June 2017
Available online xxx

Keywords:
Sketch based image retrieval (SBIR)
Deep learning
Triplet loss function
Cross-domain modelling
Compact feature representations

ABSTRACT

We present an efficient representation for sketch based image retrieval (SBIR) derived from a triplet loss convolutional neural network (CNN). We treat SBIR as a cross-domain modelling problem, in which a depiction invariant embedding of sketch and photo data is learned by regression over a siamese CNN architecture with half-shared weights and modified triplet loss function. Uniquely, we demonstrate the ability of our learned image descriptor to generalise beyond the categories of object present in our training data, forming a basis for general cross-category SBIR. We explore appropriate strategies for training, and for deriving a compact image descriptor from the learned representation suitable for indexing data on resource constrained e. g. mobile devices. We show the learned descriptors to outperform state of the art SBIR on the defacto standard Flickr15k dataset using a significantly more compact (56 bits per image, i. e. ≈ 105KB total) search index than previous methods. Datasets and models are available from the CVSSP datasets server at www.cvssp.org.

1. Introduction

The Internet is transforming into a visual content distribution network, with Cisco predicting that by 2018 over 80% of traffic will comprise visual media — two-thirds of which will be consumed on mobile devices (Cicso, 2014). Gestural interfaces, such as sketch, provide a intuitive method for interacting with visual content on such touch-screen devices, motivating the development of sketch based image retrieval (SBIR) algorithms to make sense of this deluge of visual content.

The principal contribution of this paper is to explore SBIR from the perspective of a cross-domain modelling problem, in which a low dimensional embedding is learned between the space of sketches and photographs. The embedding is learned via regression using a convolutional neural network (CNN) with triplet architecture. Such networks extend the idea of distance metric learning (Weinberger and Saul, 2009), and are trained by exposure to exemplar triplets comprising, in our case, a sketch accompanied by both a hard positive and negative example of a photograph corresponding to that sketch. Convergence is encouraged using a non-classical triplet loss function that encourages sketches and photographs of the same object to map to similar locations in the learned embedding. Once trained, the branches of the network corresponding to sketch and image input can be used to encode arbitrary images (in a search corpus) and query sketches respectively, yielding real-valued descriptors than may be compared using standard metrics (e. g. Euclidean distance) for the purpose of ranking images according to their relevance to a sketch at query time.

Deeply learned triplet ranking models have been traditionally applied to face recognition (Chopra et al., 2005; Hadsell et al., 2006) and to cross-domain problems including 3D pose estimation (mapping between images and pose space) and scene description (mapping between images and natural language spaces) (Vinyals et al., 2015). In such problems where domains vary significantly in dimensionality and statistical distribution, it is common for branch weights of the CNN to be optimised independently (heterogeneous or 'unshared' weights). Recently and most relevant to this work, triplet ranking models have been explored for visual search of images, both for photographic queries by Wang et al. (2014) and for sketched queries by Yu et al. (2016). In these works the architecture of each network branch is not only homogeneous, but also fully shares weights with its peers (i. e. siamese design) reflecting an assumption that a single learned function may optimally map the two distinct input domains to a single metric embedding. Consequently a weakness of these two prior works is a restriction to so called 'fine-grained' retrieval, in which search is constrained to operate over only a single category of object experienced during

* Corresponding author.
E-mail addresses: t.bui@surrey.ac.uk (T. Bui), leonardo.sampaio.ribeiro@usp.br (L. Ribeiro), ponti@usp.br (M. Ponti), j.collomosse@surrey.ac.uk (J. Collomosse).

http://dx.doi.org/10.1016/j.cviu.2017.06.007
1077-3142/

network training. Whilst this is effective in refining search with a dataset containing a specific object class (e. g. a dataset of shoes (Yu et al., 2016)) the approach cannot generalise to more diverse datasets, such as Internet images 'in the wild' of arbitrary subject. Such variations are exhibited in current SBIR benchmarks e. g. Flickr15k (Hu and Collomosse, 2013) which contain multiple object classes, as well as exhibiting variation within each object class. We address this limitation in our work by both adopting a partially shared weighting strategy (only the deeper layers of our network branches share weights) and modified triplet loss function. We study the ability of our network to generalise beyond the corpus of object classes exposed during training to learn an embedding for measuring SBIR visual similarity that is capable of generalising across categories.

A further contribution of this paper is to report on strategies for compactly representing the descriptors extracted from our learned embedding, to enable efficient indexing of the dataset on resource constrained platforms such as mobile devices. Compact representation of image descriptors has received considerable attention over the past five years, largely focused upon shallow-learned Bag of Visual Words (BoVW), VLAD (Jegou et al., 2009) or Fisher representations (Perronnin et al., 2010). Such approaches often combine dimensionality reduction with binary hashing (Jegou et al., 2009; 2010), for example PCA (or even, counter-intuitively, random) projection and binarisation on sign and Product Quantisation (Jegou et al., 2011). For the first time, we explore such strategies for SBIR showing that typical approaches do not operate optimally on our representation and make recommendations for accuracy preserving compaction of our descriptors. Ultimately we show that a twist on Jegou et al. 's PCA method (Jegou et al., 2009) resulting in an encoding of less than one word (i. e. 56 bits) per image is sufficient to yield an accuracy level exceeding state of the art SBIR systems with a query time in the order of 10^{-3} s for a dataset in order of 10^4 images without indexing tricks (i. e. single thread, exhaustive linear search). In practical terms, this leads to a mobile storage footprint for a contemporary SBIR benchmark dataset of 15k images in the order of one hundred kilobytes (Flickr15k (Hu and Collomosse, 2013)).

The remainder of this paper is organised as follows. We review related sketch based visual search literature, focusing upon image retrieval, in Section 2. We present our network design, configuration and training methodology in Sections 3.1–3.3 and compaction strategy in Section 3.4. We evaluate the performance of our SBIR algorithm both in terms of accuracy and storage footprint in Section 4, concluding in Section 5.

2. Related work

Sketch based image retrieval (SBIR) began to gain momentum in the early nineties with the color-blob based query systems of Flickner et al. (1995) and Kato (1992) that matched coarse attributes of color, shape and texture using region adjacency graphs defined over local image structures (blobs). More efficient matching schemes for blob based queries, using spectral descriptors (e. g. Haar Wavelets (Jacobs et al., 1995) or Short-Time Fourier Transform (Sciascio et al., 1999)) were subsequently proposed. This early wave of SBIR systems was complemented in the late nineties by several algorithms that accept line-art sketches, more closely resembling the free-hand sketches casually generated by lay users in the act of sketching a throw-away query. Since line-art depictions of objects are often stereotyped and subject to non-linear deformation (Collomosse et al., 2008), robust edge matching algorithms are required deliver usable results. Early approaches adopted optimisation strategies that sought to deform a model of the sketched query to fit candidate images, ranking these according of their support for the sketch. Models explored by these approaches include elastic contour (Bimbo and Pala, 1997) and super-pixel aggregation (Collomosse et al., 2009). However, model fitting approaches scale at best linearly with dataset size with each comparison often taking seconds leading to poor scalability.

SBIR enjoyed a resurgence of interest in the late 2000s yielding several approaches to extract real-valued feature descriptors for matching spatial structures between sketches and photograph. Eitz et al. (2009) partitioned the image into a spatial grid, computing and concatenating the structure tensor from each grid cell to form a efficient index representation, scalable to thousands of images with interactive speeds. Gradient domain features (Dalal and Triggs, 2005) were further exploited by Hu et al. (2010) who transferred the popular 'bag of visual words' (BoVW) paradigm to SBIR improving scalability and affording greater robustness to affine deformation. Notably this work introduced the Gradient Field Histogram of Gradient (GF-HoG) descriptor for encoding spatial structure within the BoVW representation using Laplacian smoothness constraint. GF-HOG was later robustified through extension to multiple scales (Hu and Collomosse, 2013) and evaluated over the Flickr15k benchmark which remains the largest fully annotated dataset for SBIR evaluation. The BoVW framework of Eitz et al. alternatively proposed sparse S-HOG and SPARC descriptors (Eitz et al., 2011) for encoding local edge structure in the codebook, which were later combined with SVM classification for sketch based object recognition (Eitz et al., 2012), yielding also the first large scale dataset of sketched objects (TU-Berlin). Qi et al. demonstrated that perceptual edge extraction during pre-processing can boost performance of HOG-BoVW pipelines, maintaining the state of the art position on Flickr15k of 18% mAP (Qi et al., 2015) but with query times of minutes. The GF-HoG framework was extended to multiple modalities (including motion) by James and Collomosse (2014) and a detailed study of its extension to color SBIR was made by Bui and Collomosse (2015), who also adapted the technique of inverse-indexing to SBIR for large-scale search (3 million images in a few hundred milliseconds). Edgel based index representations for SBIR were explored by Cao et al. 's MindFinder (Cao et al., 2011) and were exploited in the assisted sketching interface ShadowDraw (Lee et al., 2011). A mid-level representation combining edgels into 'key shapes' was also explored by Saavedra and Bustos (2014) and Saavedra and Barrios (2015).

Deep convolutional neural networks (CNNs) have revolutionised Computer Vision, providing a unified framework for learning descriptor extraction and classification that yield major performance gains over pipelines based upon prescriptive gradient features (Krizhevsky et al., 2012). CNNs have similarly impacted sketch. The Sketch-A-Net CNN architecture introduced by Yu et al. (2015) improved upon the accuracy of Eitz et al. 's BoVW-SVM framework (Eitz et al., 2012), mirroring human level performance at object categorisation over TU-Berlin. Similar to the architecture of our proposed system, Yu et al. (2016) very recently adapted Sketch-A-Net into a triplet loss siamese architecture to effect SBIR within a single object class (e. g. shoe search). In this paper we use a deep convolutional network that uses a triplet loss function in order to learn a cross-domain mapping between sketches and edge maps obtained from natural images (Arbelaez et al., 2011), capable of generalising SBIR across hundreds of object categories. Several aspects of our design are novel to SBIR; unlike the multiple stage training process of Li et al. (requiring independent training, fine-tuning and integration of branches of the siamese network) we learn a depiction invariant representation as a single regression across the triplet CNN. We introduce a partial weight sharing scheme and novel triplet loss function to encourage convergence when learning our cross-domain embedding; a difficulty also observed in Wohlhart and Lepetit (2015) inherent to cross-domain matching. Further, we show that our partial weight sharing scheme – in effect, the learning of separate functions for

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

each domain – outperforms fully shared (siamese) strategies, both in triplet and contrastive loss (Qi et al., 2016) networks as very recently proposed.

Previous studies using of triplet networks for cross-domain matching include FaceNet (Schroff et al., 2015) which learns a mapping from face images to a space in which distances directly correspond to a measure of face similarity. A similar method was used also by Wang et al. (2014) to learn a ranking function for image retrieval. The triplet loss concept has more widely been used for cross-domain mapping between text (skip-gram features (Mikolov et al., 2013)) and images (Frome et al., 2013) and in the context of RGB-D 3D data (Wohlhart and Lepetit, 2015), in order to learn a descriptor that includes both the 3D object identification and its pose, allowing to classify the object and its pose at the same time. In the latter authors evaluated the method in a dataset with 15 classes and accurate ground truth for the training stage. Recently, a magnet loss was proposed by Rippel et al. (2016) to consider multiple examples at a time for the classification task with focus on the fine-grained classification, which does not suit the SBIR task. Siamese triplet networks have been applied for video classification, learning from unlabelled videos by finding positive pairs through object tracking (Wang and Gupta, 2015).

This paper also explores efficient representations for encoded the learned embedding for SBIR. Compact feature descriptors have been extensively studied for large-scale visual search using photographic queries where feature quantisation via binarization (Jegou et al., 2009) and either PCA or random basis projection (Jegou et al., 2010; Perronnin et al., 2010), as well as Product Quantisation (PQ) (Jegou et al., 2011). We compare and contrast these approaches for our proposed SBIR representation, and propose an adaptation of one for efficient search on resource constrained devices such as mobile tablets.

3. Learning a compact representation for SBIR

We approach SBIR as a regression problem, in which an embedding is learned via a triplet neural network with partially shared weights using a modified loss function that promotes better convergence of the network. We first outline the architecture of our network in Section 3.1 and describe the loss function in 3.3. We then outline the network training process in Section 3.2, including an account of the data acquisition and pre-processing necessary to perform this training. In the final step of our process a quantisation based on PCA projection is performed to obtain a compact representation, discussed in Section 3.4.

3.1. Triplet network architecture

Single-branch discriminative networks CNNs (e. g. AlexNet (Krizhevsky et al., 2012)) have delivered transformational performance gains in visual concept detection. These successes have been mirrored in the sketch domain, notably in (Yu et al., 2015) work where the Sketch-A-Net CNN was shown to parallel human performance on sketched object recognition over the TU-Berlin dataset. Recent work indicates that more modest gains can be obtained through fusion of classifiers (e. g. SVMs) with features extracted in fully connected (fc7) network layers (Razavian et al., 2014). One might naïvely hope that such layers might also provide learnable features for visual search. However such approaches generalise poorly to imagery beyond the training set domain, motivating many to employ triplet CNN network designs (Wang et al., 2014; Wang and Gupta, 2015). Rather than forcing all input data to converge to a single point in the CNN output space (as implied in a softmax loss), the triplet loss constrains only relative distance between the anchor, and the positive and negative exemplars (Fig. 2a and b). In our case the triplet comprises a sketched query for the anchor branch, and positive and negative exemplars derived from edge maps of photographs (Section 3.2.3).

In contrast to prior work (Li et al., 2015; Wang et al., 2014) that incorporates three identical and siamese CNN branches into their triplet networks, we propose a "half-sharing" scheme across network branches which we argue better fits the problem of general SBIR (Fig. 1, left). The network design comprises an adapted form of Yi et al. 's Sketch-A-Net, duplicated across the three network branches and unified in a final fully connected layer of low dimensionality (we use $\mathfrak{R}^{100}$ for all experiments reported) to encode the embedding. The positive and negative branches that accept photo images are identical, whilst the anchor branch accepting sketched input shares only the 4 top layers with the other branches. Whilst similar in dimensionality, sketch and image edge-maps encode very different domains, thus their visual feature extraction frameworks (encapsulated within the first few layers of the CNN) should not be the same. This assumption is borne out in Fig. 1 (right), where the first layer of 64 filters has been visualised from the sketch and photo branches of the CNN showing clear differences between the pair of domains. Although we explored configuring the anchor branch to be completely independent of the other two branches, our experiments conclude (c.f. Section 4 that over-relaxing the CNN model in this manner easily leads to over-fitting.

3.2. Training and data acquisition

We learn the sketch-photo embedded by training our CNN using exemplar triplets comprising a free-hand sketch (the anchor), and two photographs corresponding to a positive search results (i. e. matching the object class of the sketch) and a negative search result (i. e. any other object class). We now outline this process in detail.

3.2.1. Datasets

We used two public sketch datasets in order to run our experiments: the **TU-Berlin** classification dataset (Eitz et al., 2011) and the SBIR **Flickr15K** dataset for visual search (Hu and Collomosse, 2013):

- The TU-Berlin classification dataset, used in the **training** stage only, comprises 250 categories of sketches and 80 sketches per category. Because these were crowd-sourced from 1350 different non-expert subjects, the drawing styles are diverse. In addition to the sketches available in this dataset, our method requires photographs images to match the sketch categories. Those images were obtained from the Internet as described in Section 3.2.2.
- The Flickr15K, used for **testing**, has labels for 33 categories of sketches and 10 sketches per category. It also has a different number of real images per category totalling 15,024 images. The Flickr15K dataset has few shared categories with the TU-Berlin dataset (only "swan", "flowers", "bicycle" and "airplane"). The two datasets also conflict in a few categories, for example TU-Berlin has a general "bridge" category while Flickr15K distinguishes "London bridge", "Oxford bridge" and "Sydney bridge". In addition to the limited category overlap, the datasets differ in their sketch depiction (only a few commonalities exist in objects with simple structure, such as circles depicting the moon). This challenge motivates a need for good generalisation beyond training.

3.2.2. Internet photo acquisition

To generate the training triplets we automatically generated sets of photographs corresponding to each TU-Berlin object category by downloaded content from the Internet. For each category,

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

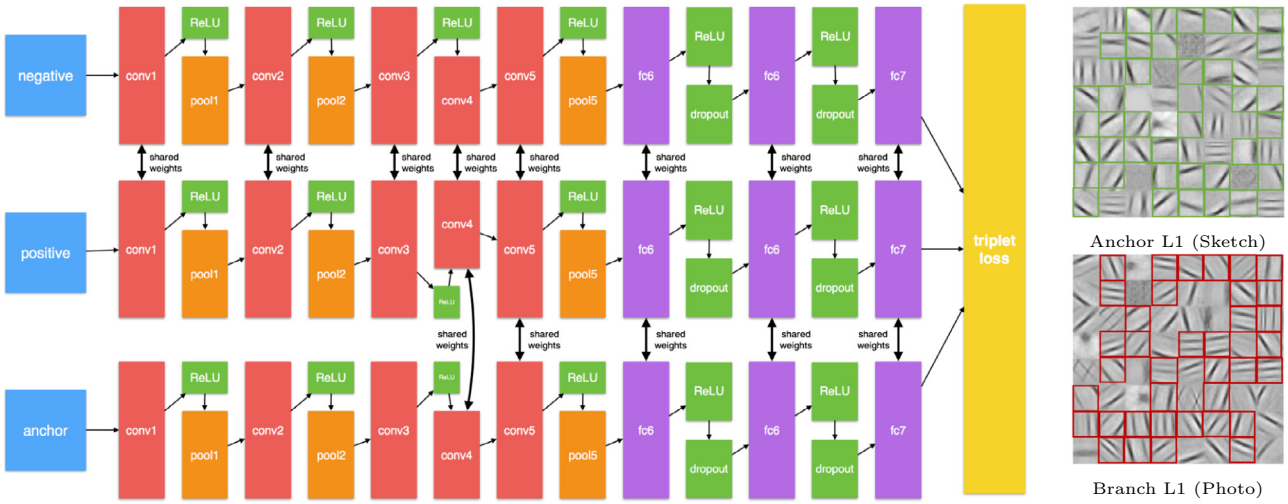


Fig. 1. Left: The proposed "half-sharing" triplet CNN architecture for SBIR. The query sketch is fed to the anchor network while a pair of photographic images are directed to the right positive/negative nodes. Right: Visualisation of the 64 filters within the first layer of the anchor (sketch) and shared branches (photograph). The green and red boxes highlight single and double edge filters. These are highlighted within the sketch and photo edgemap branches respectively — the composition of the filter banks of each branch are quite different, supporting independent learning of the early stage layers in our half-share framework. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

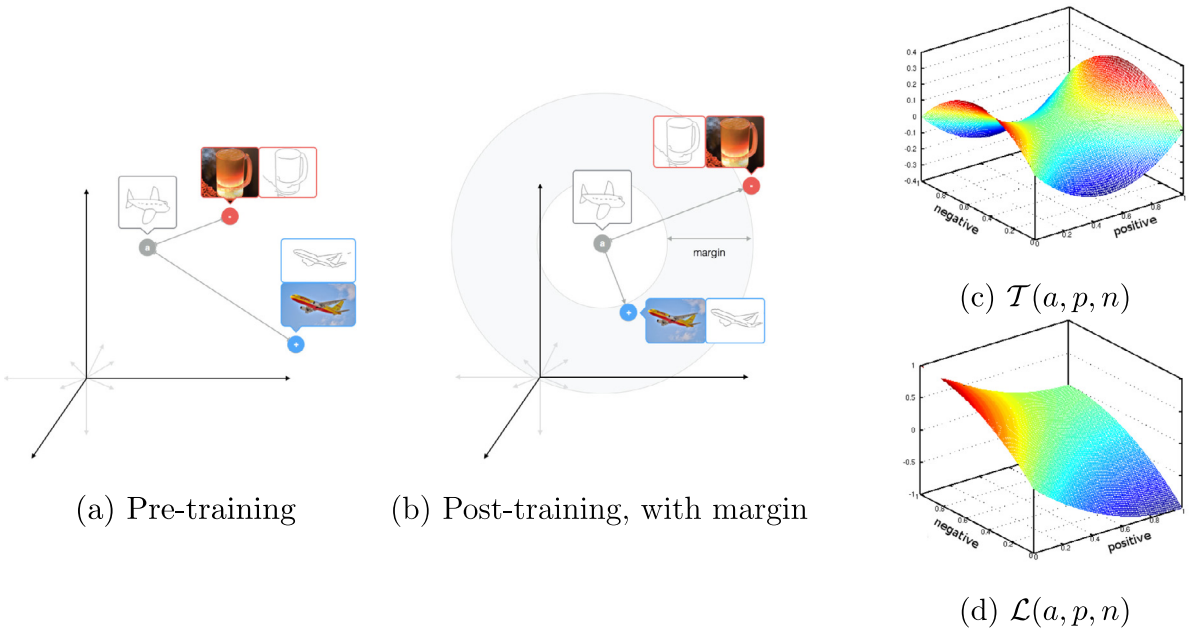


Fig. 2. Left (a-b): Triplet loss minimises the distance between an anchor and a positive exemplar (same category) and maximises the distance between an anchor and a negative exemplar (any other category). Right: Comparing (c) conventional triplet loss function, and (d) proposed weighed-anchor triplet loss function for a fixed value of $a = 0.5$.

we queried the Flickr API using the category name as the search term, limiting results to the most relevant 100 images on the basis that larger result sets were frequently polluted with non-relevant images. For a few unusual categories that are unpopular on Flickr (e. g. human brain, tooth, human skeleton, ...) we downloaded images using the Bing and Google Images APIs. Specifically, the Flickr API was used to download images from 184 categories, with the search engines covering the remainder of content.

3.2.3. Image pre-processing

Prior to using the harvested photos as input to CNN training, the images are pre-processed. First, images are resized maintaining aspect ratio such that longest side (height or width) of the image is 256 pixels. This may introduce white space into the image. Second, a 'soft' edge map is computed from the resized image using the state of the art gPb contour detection method (Arbelaez et al., 2011). The soft edge map is converted into a binary edge map by thresholding to retain the 25% strongest edge pixels, and removing the weakest 25%. Edge following (hysteresis thresholding) is applied to the remaining pixels as per the Canny edge detection algorithm to detect the edge points connected to the "strong edges" and remove the isolated edge pixels. The resulting edgemap is padded with non-edge pixels to achieve a fixed dimension of 256 × 256 pixels.

Prior to using the TU-Berlin sketches as input to the CNN training, skeletonisation is performed (Lam et al., 1992). As an artefact of the capture technology used in TU-Berlin, sketches in that dataset consist of thick strokes whereas edge maps obtained from

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

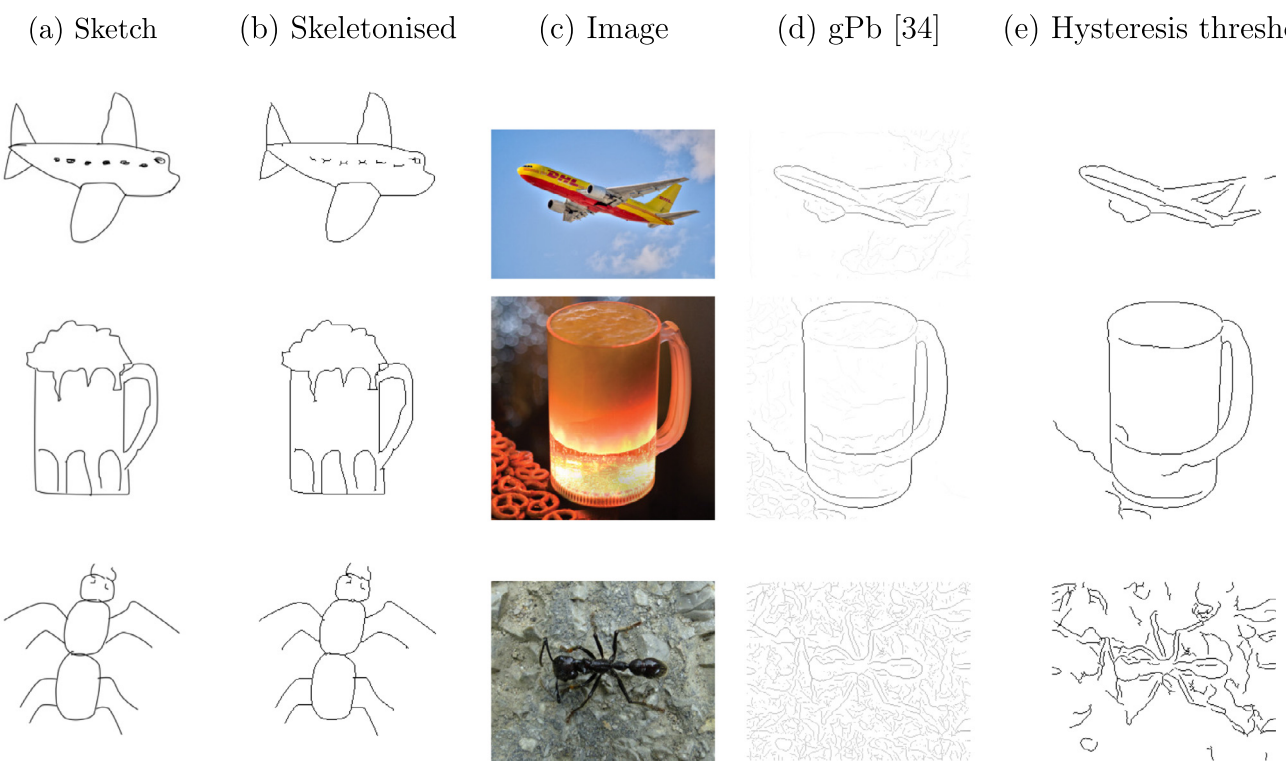


Fig. 3. Pre-processing of images for triplet exemplars. Insets (c–e) in the last row are a failure case of gPb (Arbelaez et al., 2011) due to strong edges detected in background.

the images exhibit pixel-thin contours. The skeletonisation process brings the two domains closer into alignment.

Fig. 3 summarises the image and sketch pre-processing steps applied.

3.2.4. Triplet selection and training

We train the CNN for 10,000 epochs using stochastic gradient descent (SGD) implemented in PyCaffe (Jia et al., 2014), step-decreasing the learning rate at epoch 6000, 8000 and 9000. During training, a triplet is formed by randomly matching a sketch with an image's edge map of the same class as a positive pair, and a random image's edge map of a different class as a negative pair. We use a batch-size of 200 triplets, each containing 20 classes sampled at random from those available in the data (up to 250 classes, as per TU-Berlin). Although randomly selected, we ensure even sampling across the 20 classes within each mini-batch (we found this detail important to promote convergence).

3.3. Weighed-anchor triplet loss function

The conventional triplet loss function is defined as:

$$\mathcal{T}(a,p,n)=\frac{1}{2N}\sum_{i=1}^{N}\max\left[0,m+|a_i-p_i|^2-|a_i-n_i|^2\right], \tag{1}$$

where m is the distance margin. However, in our experiments with large number of categories, the training quickly converges to a global loss of $m/2$. This typically occurs when the differences $|a-p|$ and $|a-n|$ produced low values at the same time for a given input triplet. Instead of learning to pull the positive image closer to the anchor, whilst pushing the negative image away, the network effectively draws all 3 points closer together in the learned embedding. Consequently, the gradient with respect to the three inputs collapses to zero and the network stops learning (Fig. 5a).

A common tactic to discourage gradient collapse is to consider L2 normalisation of the output layer prior to evaluation of Eq. (1) i. e. to enforce the two domains to lie upon the unit sphere. However, in this case we are not dealing with loss spiking and collapsing to zero, but rather to oscillate then flatten to $m/2$ (Fig. 5a). Wohlhart and Lepetit (2015) observed similar difficulties in their own cross-domain learning, and proposed a new triplet loss function where the margin and anchor-negative distance are moved to a denominator of the loss function. However, we found this function to be unstable for our sketch-photo mapping. Moreover we found that manual initialisation with several margin values or tuning of the learning rate, batch-size and filter parameters was unhelpful. We instead propose a modification of the conventional loss function in order to aid convergence by weighing the anchor vector with a k factor as follows:

$$\mathcal{L}(a,p,n)=\frac{1}{2N}\sum_{i=1}^{N}\max\left[0,m+|ka_i-p_i|^2-|ka_i-n_i|^2\right]. \tag{2}$$

The gradient for the backpropagation is given by the derivatives of this weighed-anchor loss function for positive loss with respect to each term:

$$\frac{\partial\mathcal{L}}{\partial a}=-\frac{k}{N}[p-n],$$

$$\frac{\partial\mathcal{L}}{\partial p}=\frac{1}{N}[p-ka], \tag{3}$$

$$\frac{\partial\mathcal{L}}{\partial n}=-\frac{1}{N}[n-ka].$$

Fig. 2c and d visualises the two loss functions $\mathcal{T}(a,p,n)$ and $\mathcal{L}(a,p,n)$ varying $(p,\ n)$ for a fixed value of $a=0.5$ and showing that $\mathcal{T}(a,p,n)$ is prone to developing a saddle point. This example considers the anchor, positive and negative as scalars analogous to the scalars generated when computing the differences in the latter two terms of each loss function. The plots indicate that relative differences between the positive and negative vectors can develop a surface exhibits a saddle, in which convergence is hampered. By

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

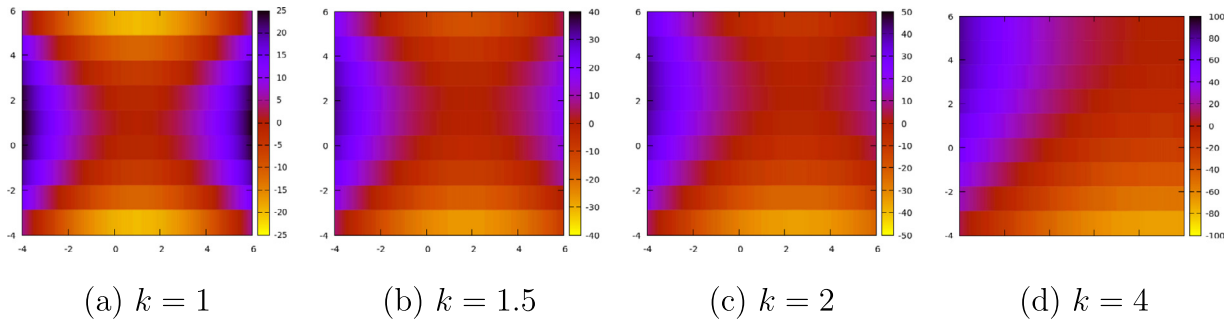


Fig. 4. Heatmaps showing a fixed search region distorted by increasing the weight of the anchor on the proposed loss function. Here the anchor vector a and the axis represent different relative values for p and n.

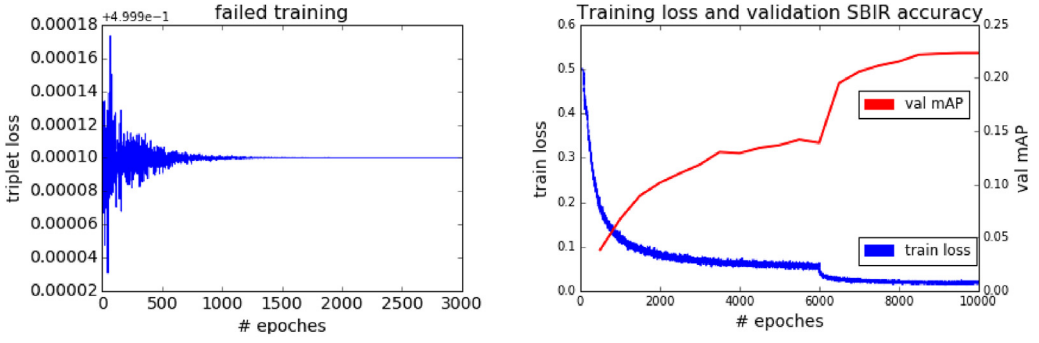


Fig. 5. (a) Illustrating collapse of the gradient with the classical triplet loss, and (b) training with our new loss function. Note the effects of decreasing the learning rate at epoch 6000. At epoch 8000 and 9000, the learning rate is already too small to have a significant impact. Plotting training loss (train loss) and validation mAP (val mAP).

increasing the weight of the anchor via the parameter k, the function is warped, yielding a surface with more defined slope, so aiding convergence of the optimisation and mitigating the problem. This effect is visualised using real triplet data in Fig. 4 for a fixed search region and different values of k. Note that for higher values of k the saddle is removed from the search region and that there is a clearer slope in the resulting surface that resembles a linear function. The SGD is now encouraged to fit the sketch output feature as k times the image feature. This behaviour can be verified empirically by examining the first non-shared layers of the trained anchor branch, as its learned convolutional weight range is around $1/k$ of the image's counterpart. At test time, when extracting a descriptor within the embedding for a query sketch, we simply multiply the output of the anchor branch with k. This must be done prior to using the anchor point as a basis for measuring Euclidean distance in the space to assess visual similarity for SBIR. We later explore the sensitivity of retrieval performance, and of convergence time to parameters m and k (Section 4.2).

3.4. Embedding dimensionality, PCA-Q and PQ

Having trained the CNN, the convolutional weights are fixed and the sketch (anchor) and photograph branches extracted to provide a pipeline for feature extraction. At test time, the query sketch and all photographs within the dataset are passed through the respective branches to obtain 100-D descriptors (for the latter this is performed once during content ingestion and indexing). In our experiments, the Euclidean (L2) Distance is used to exhaustively compare the query and image descriptors using the distance as a ranking metric for relevance. More sophisticated indexing strategies that scale sub-linearly with dataset size are commonly found in the information retrieval literature (e. g. clustered retrieval via hierarchical K-means (Nister and Stewenius, 2006) or kd-trees (Hu and Collomosse, 2013)) however such engineering is outside the scope of this paper's contribution, which focusses on the feature representation itself.

The use of a triplet architecture and the loss layer on our CNN makes the learning task more similar to a regression than a classification. Therefore the output dimensionality can be seen as a hyperparameter of the system. It was shown by Schroff et al. (2015) that the dimensionality does not affect the final accuracy in a statistically relevant manner, provided it is high enough to encompass the output space variance. With that in mind, we selected 100 as our output dimensionality, which provides an efficient training time while maintaining the capability of representing descriptors of varying complexity. To evaluate our choice two dimensionality reduction methods were used: a Product Quantisation (PQ) and a Principal Component Analysis based Quantisation (PCA-Q). In the case of PCA-Q, we have the following steps (i) compute PCA and explore dimensionality reduction using the first principal components (PCs); (ii) perform quantisation on PCs, in terms of number of bits used to represent each dimension (Ponti et al., 2016).

Practically, it is useful to have a compact representation for both retrieval and classification tasks to enable efficient storage and transmission of the search index e. g. for mobile devices. Strategies for compact representation of visual search descriptors have been explored extensively for classical (i. e. photo query) visual search. The PQ-based approach is able to obtain compression but the space of parameters is large, and therefore it is harder to find the optimal values, as we show in the following section. In addition to PQ, we investigated a feature compressing method by performing PCA projection of the features as in Sun et al. (2014), also exploring subsequent quantisation of each dimension into a n-bit fixed point integer. For our SBIR representation we found that compaction without significant loss of accuracy can be achieved

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

via this combination of PCA projection and independent quantisation of descriptor elements.

4. Experiments and results

We report experiments exploring both the accuracy and storage overhead of an evaluation SBIR system constructed over our learned representation base upon an exhaustive linear comparison of the dataset using Euclidean distance. We begin by describing the experimental setup and significant parameters to performance in Section 4.1. We explore the accuracy of the learned representation in Section 4.2 without any compaction. In particular we characterise our system's ability to generalise to category-level SBIR in the Flickr15k dataset, from a limited and distinct set of object categories observed during training. In Section 4.3 we compare our descriptor compaction approach to classical approaches, exploring the effect of storage overhead and accuracy.

4.1. Datasets and experimental settings

We trained our network on a single NVidia Titan-X GPU hosted within an Intel i7 5930K CPU Server with 128GB RAM. Having trained the embedding, we run our visual search experiments on an Intel Xeon E5-2643 CPU Server with 200 GB RAM. It is on the latter machine that our timings in (Section 4.3) are reported. All training was performed on the TU-Berlin dataset augmented with creative-commons Internet imagery, as outlined in Section 3.2.1. During the test phase the Flickr15k SBIR benchmark was used to obtain mAP and precision-recall statistics, as this was the largest available SBIR dataset of fully annotated images (15,024 photographs) for 330 individual sketch queries (Hu and Collomosse, 2013).

Unless otherwise noted, we fixed the triplet margin value $m = 1.0$ and loss scale factor $k = 3$ (Eq. (2)) for all experiments reported here. Training time depended on the experiment, for instance typical time to train a 250-category triplet model for 10k epochs with batch size 200 was 6 days. That is less than a minute for one pass through the sketch (anchor) training dataset. We used SGD for optimisation in all experiments. We used a multi-step learning rate during training, initially set to 10^{-2}; Fig. 5 provides a representative example of the effect of learning rate on convergence.

In order to compensate for the limited sketch data available for training (a difficulty for all contemporary sketch CNN research) and to combat the over-fitting problem, we apply data augmentation. In addition to standard mean-image (here, mean-edge) subtraction, sketch and photograph edge maps are randomly cropped, mirrored and rotated within a range of $[-5,5]$ degrees. Additionally, we created new sketches by randomly discard strokes from the original SVG sketches in TU-Berlin dataset, considering drawing order as noted in Yu et al. (2015). The augmentation process was implemented on-the-fly via a custom data-layer to reduce storage space.

4.2. Evaluation of generalisation

A series of experiments were conducted to investigate the ability of the triplet ranking model to generalise beyond its training data for category-level SBIR. We constructed training sets with different number of categories: 20, 40, 80, 130 and 250 sampled arbitrarily at random from the TU-Berlin dataset. We also varied number of training sketches per category: 20, 40, 60 and the whole 80 sketches/category were randomly sampled for training. Additionally, we experimented with different sharing levels between sketch and edge-map branches. We start from sharing the whole branch (full-share), then unlock sharing each layer from bottom to top and finally not sharing any layer at all. Having trained the CNN, mAP over the Flickr15k test dataset was evaluated, in order to measure how well the embedding was able to generalise to unseen categories.

Table 1
SBIR results comparison (mAP) against state-of-the-art alternatives.

Methods	Feature size (bits)	mAP (%)
Proposed approach	3200	**24.45**
Proposed approach (compact)	**56**	22.03
Contrastive loss (Qi et al., 2016)	2048	19.54
PeceptualEdge (Qi et al., 2015)	–	18.37
Triplet full-share (Yu et al., 2016)	3200	18.36
Color gradient (Bui and Collomosse, 2015)	160,000	18.20
PeceptualEdge-proximity (Qi et al., 2015)	–	16.02
GF-HOG (Hu and Collomosse, 2013)	112,000	12.22
HOG (Dalal and Triggs, 2005)	96,000	19.93
SIFT (Lowe, 2004)	32,000	9.11
SSIM (Shechtman and Irani, 2007)	16,000	9.57
ShapeContext (Belongie et al., 2002)	112,000	8.14
StructureTensor (Eitz et al., 2009)	16,000	7.98
PeceptualEdge-continuity (Qi et al., 2015)	–	7.89

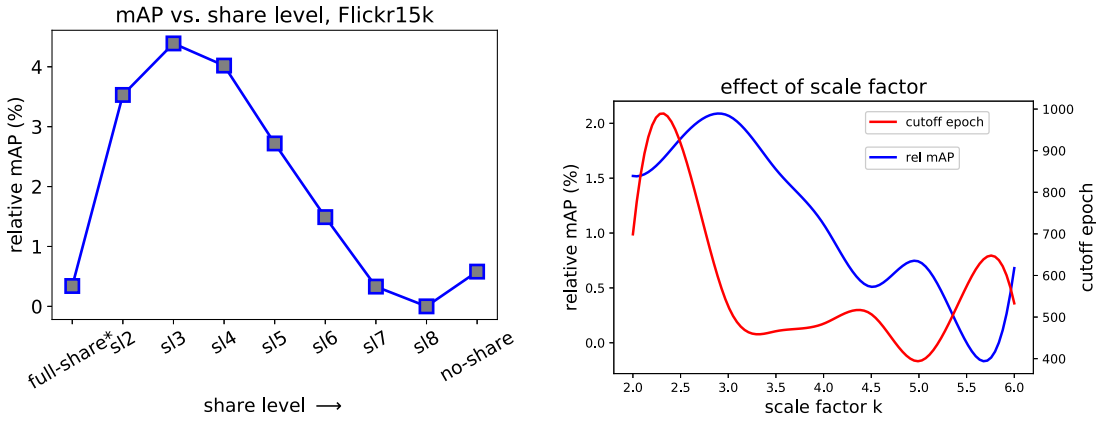
We successfully trained the triplet networks with loss Eq. (2). The imbalance in the loss function enables the sketch branch to correctly bias the other branches helping to overcome the collapsing gradient problem (see Section 3.3) and the training loss decreases rapidly after the first few hundred epochs (Fig. 5). Fig. 6a explores the effect of varying the sharing policy between branches during training. We observe that partially sharing the weights yields superior performance over the full-share and no-share configurations. A gain of $\approx 3\%$ mAP is obtained after the first convolutional layer is released from sharing, allowing the bottom layers to learn distinct functions for their respective domains. The best half-share design is performance is obtained when sharing from layer 3.

We also evaluated the effect of scale factor k in Eq. (2) (Fig. 6b) in term of performance (mAP) and learning speed (measured by cut-off epoch, which is the epoch number where the training loss avalanches to $1/\sqrt{2}$ of the initial loss, or $m/(2\sqrt{2})$. Within the working range [2.0, 6.0], best performance and convergence speed was achieved at $k = 3.0$. The training does not converge beyond $k < 2.0$ and $k > 6.0$. We experimented within a working range of margin $m = [0.5, 1.5]$ observing negligible change in mAP or convergence time. Beyond this range, a small margin leads to an easy task where the training converges early to a suboptimal solution (the triplet condition is satisfied too easily). In contrast, a large margin inhibits training convergence.

Fig. 7b and c illustrates the near-linear improvement of mAP as the category count and sketch-per-category count increase. Overall, we achieve $\approx 2.5\times$ increase in retrieval precision when expanding the training database from 20 to 250 categories. At 250 categories, a boost of 2–4% mAP was achieved when increasing number of training sketches per category from 20 to 80. Fig. 8 visualises the distribution of the first 6 categories in the Flickr15k dataset whose embedding features are extracted from 20 and 250 category trained models respectively. A qualitative improvement in inter-class separation is observable using the larger number of training categories, mirroring the performance gains observed in mAP as category count increases. Quantitatively after increasing the number of training categories from 20 to 250, the inter-class distance of the embedding features is pushed further away by 33%, while the average intra-class distance is reduced by 10%.

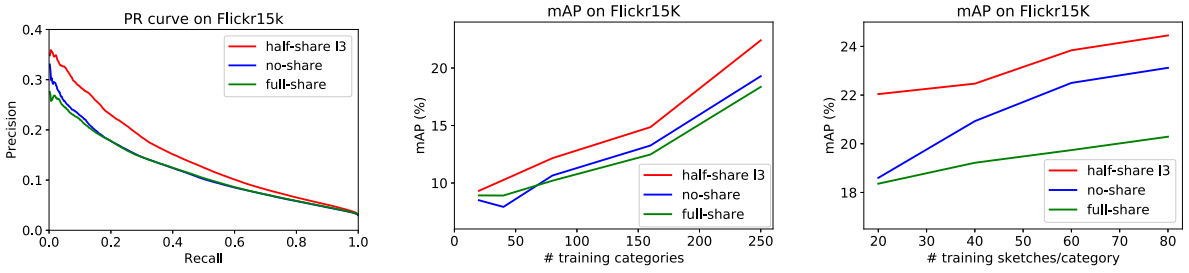
Table 1 compares our approach with contemporary SBIR algorithms, over the Flickr15k dataset. We compare our proposed approach and its compact version (Section 4.3) with two recent deep-learning baselines (Qi et al., 2016; Yu et al., 2015) (note: our triplet full-share version is analogous to Yu et al. (2016)) and 10 other approaches using shallow features. We exceed state of the art

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007



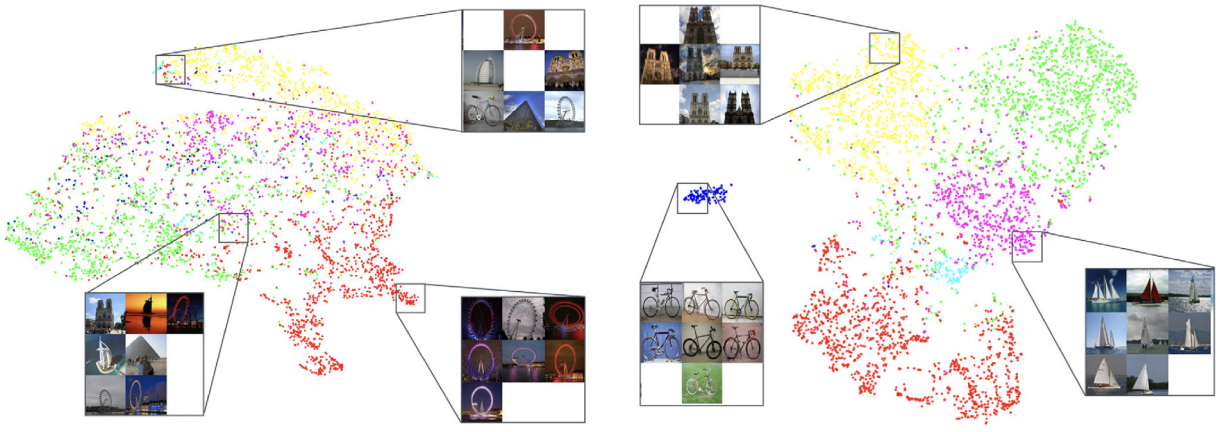
(a) level of sharing (b) Scale factor.

Fig. 6. Exploring training meta-parameters. (a) Plotting the performance impact (mAP) for various degrees of weight sharing between branches, from full-share to no-share. (b) Plotting the effect of scale factor k on performance (mAP) and time to training convergence (epochs) under Eq. (2).



(a) PR curve (b) # training categories (c) # sketches per category

Fig. 7. Relative performance (mAP) of a fully shared, no share, and partially shared (from layer 3) network: (a) Precision-Recall (PR) curve showing our state of the art result; (b) The generalisation capability of our model across test categories as number of the training categories increases; (c) the impact of increasing the number of training sketches.



(a) 20-class training (b) 250-class training .

Fig. 8. T-SNe visualisation of the Flickr15k dataset's data distribution, projected on the 3 major Eigen vectors, viewed at the same angle, for the training with (a) 20 classes and (b) 250 classes. Each color represents a semantic class in Flickrk15k.

accuracy on the Flickr15K dataset by 5%. The precision-recall (PR) curve is depicted in Fig. 7(a).

We observe little performance difference when searching with sketches of categories within the semantic overlap between Flickr15k and TU-Berlin, and those non-overlapping (the majority of Flickr15k categories). Fig. 9 shows the retrieval results for several query examples. Example queries from categories "bicycle" and "horse" (shared by both TU-Berlin and Flickr15K datasets), together with "Arc de Triomphe" and "Eiffel tower" (unseen categories), "London bridge" and "Sydney bridge" (related to the conflicting "bridge" category under the TU-Berlin dataset) return excellent results. At the same time, a query from the shared category "tree" (row second from bottom) is among the worst results. This provides further evidence for generalisation in the trained model.

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

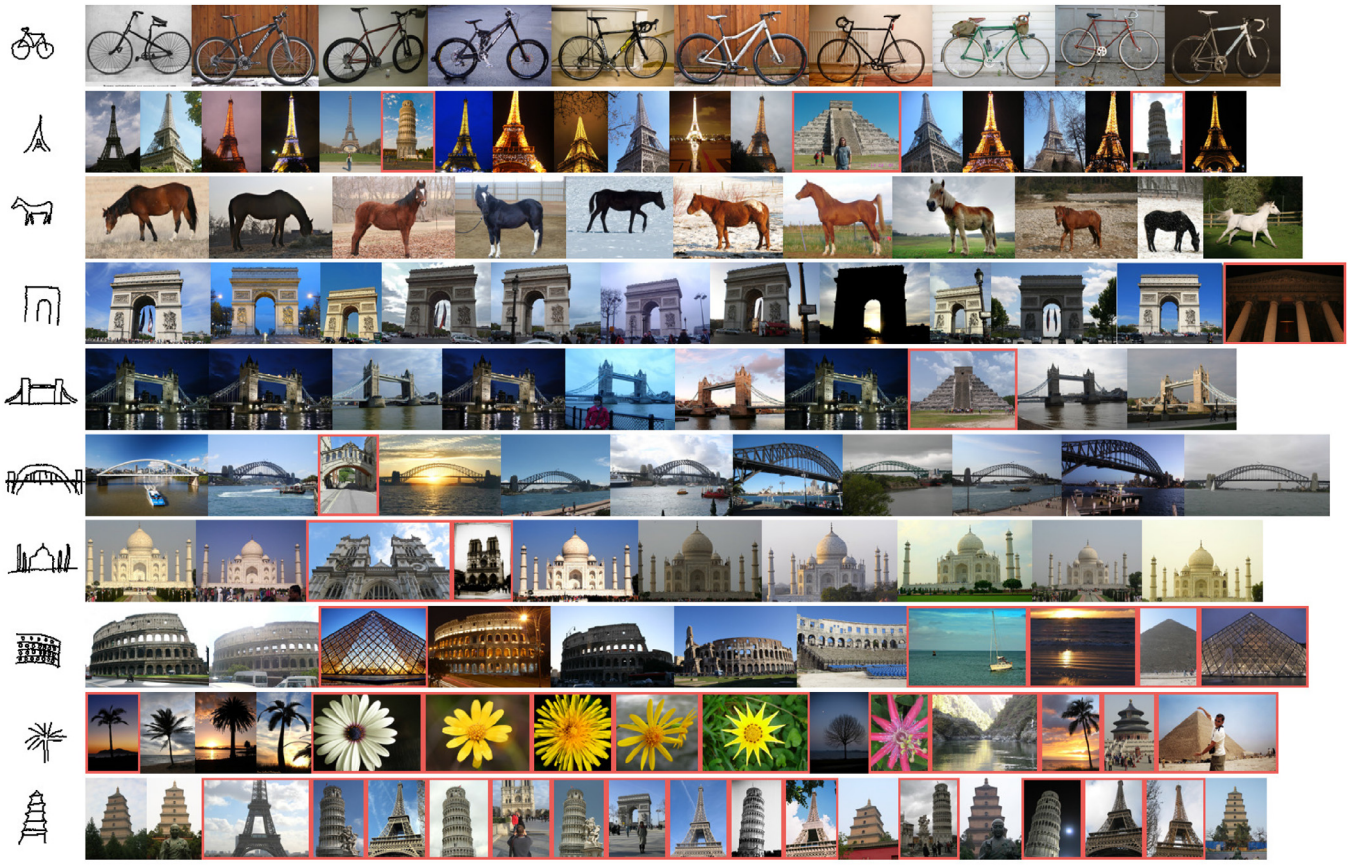


Fig. 9. Top retrieval results for several query examples. Top rows are examples of the good results while bottom rows are the bad ones. Non-relevant images are marked with a red bounding box. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3. Evaluation of compact representation

The 'raw' 100 dimensional embedding output by the triplet network yields a best-case performance when trained over 250 categories of the TU-Berlin dataset, achieving a mean average precision (mAP) of 24.45%.

Two approaches were tested in order to obtain a compact representation, the first based on Product Quantisation (PQ) (Jegou et al., 2011) and the second based on PCA. We tested PQ with various parameters for the clustering algorithm as follows: $K = 16, 32, 64, 128, 256, 512$ (coarse), $m = 2, 5, 10, 20, 30$ and $K' = 16, 32, 64, 128, 256, 512$ (fine) – these parameters are discussed fully in Jegou et al. (2011). For the PCA approach, we explore variation in the number of principal components $m = 6, 7, \cdots, 14$ and quantisation of the PCA projected values into $n = 4, 5, \cdots, 10$ bits. The relative performance of these configurations is summarised in Fig. 10.

Reducing the descriptor storage footprint further, from a floating point representation (32–64 bits) to n bits per dimension, is critical for deployment of SBIR on resource constrained devices or for scalability to a large search corpus. We explored quantisation of the real-valued space into a n bit fixed point representation for the PCA for the 250-category i. e. best case embedding. In the case of PQ and its parameters, the resulting representation requires a number of bits proportional to $1 \cdot \log_2(K) + m \cdot \log_2(K')$, while in PCA-Q the bit size is number of bits times number of principal components, i.e. $n \cdot m$.

To the best of our knowledge, no compact representations have previously been explored for SBIR. Fig. 10 (right) plots the results of PQ against our PCA-Q approach. Significant performance improvements are yielded under our proposed method i. e. PCA + byte quantisation, which is more stable than the PQ when varying the parameters for different byte sizes. We considered as optimal the use of 56 bits using PCA-Q — 14 PCs quantized to 4 bits each — since it yields 22.03% mAP (only 2% less than the full model), while the PQ result for the same 56 bits ($K = 64$, $m = 10$, $k' = 32$) is 19.52% mAP. Additionally, PCA-Q is more stable, easier to tune and implement, while PQ requires a larger parameter search. Although not significant to scalability there is also a small constant overhead of $K \cdot 64 + m \cdot k \cdot 64$ bits for the PQ codebook.

Table 2
Run-time of our proposed system for a single sketch query over the Flickr15k dataset, tabulated for multiple quantisation setups on the 250 category configuration.

PCs \nBits	4	8	16
4	14.6 ms	17.3 ms	17.9 ms
8	14.3 ms	17.5 ms	17.9 ms
12	14.0 ms	17.7 ms	18.7 ms

In addition to reduced storage overhead, compact representations for visual search are often faster to compare at query-time than real-valued representations, and this is important for deployment on resource constrained devices. Table 2 reports observed run-times (time to perform one query against Flickr15k) obtained from our quantised representation under a single-threaded Matlab implementation. Whilst the original non-quantised (floating point) descriptor the runtime of a query is **29.8 ms**, retrieval using quantised descriptor with the optimal setup (56 bit, PCA-Q) reduces the runtime by approximately **41%**. A machine-level implementation using extended instruction sets on modern CPUs (e. g. SSE3)

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

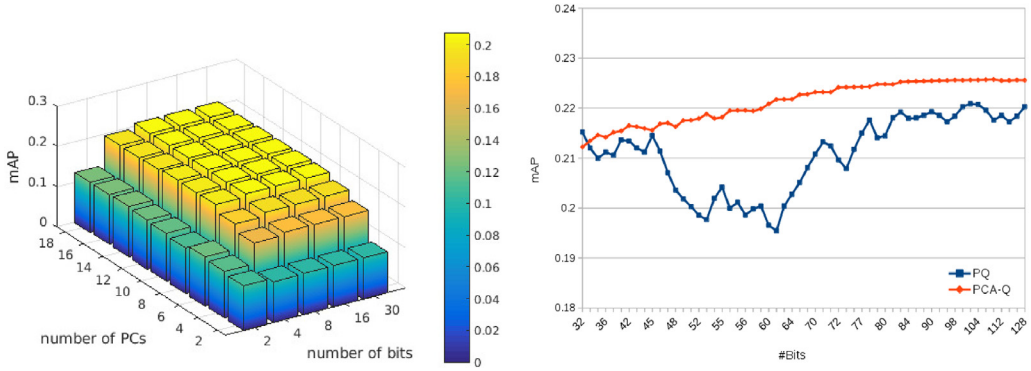


Fig. 10. Comparing performance (mAP) for different compact representation strategies. Left: Relative performance of PCA-Q for various numbers of principal components (PCs) and bit quantisations of the projected dimensions. Right: Retrieval results comparing PQ to PCA-Q highlighting that our choice of using principal components and bit quantisation is compact with consistent performance over different bit sizes. This plot was obtained by averaging chunks of 16 bits in order to smooth out noise in the parameter space. Performance levels out at 56 bits motivating selection of this bit length at the winning strategy.

could compare each image in a single instruction cycle given the size of descriptor is less than one word (64 bits).

5. Conclusions

We have presented a compact representation for sketch based image retrieval (SBIR) suitable for deployment on resource constrained mobile devices. Tackling SBIR from the perspective of a cross-domain modelling problem, we developed a triplet ranking CNN to learn a low dimensional embedding between the space of sketches and photographs. The CNN was trained using exemplar triplets comprising sketches from a public 250 object category sketch dataset (TU-Berlin) and corresponding photographs sourced from the Internet (via the FlickR API). Evaluating the trained CNN over a challenging SBIR benchmark (Flickr15k) yielded a significant performance boost to 24% mAP (or 22% in the compact representation), an improvement over previously reported results (12% (Hu and Collomosse, 2013), 18% (Qi et al., 2015), 19% (Qi et al., 2016)), using a regular distance measure (L2 norm) within the learned embedding. Notwithstanding this improvement, this work is also the first triplet ranking CNN for SBIR that is not constrained to so called 'fine grain retrieval i. e. restricted to the same single object category exposed to the CNN during training. Rather, we have demonstrated the ability to generalise beyond objects experienced during training (category level retrieval) and explored this ability over a range of training object categories. We also showed that a combination of PCA and dimension quantisation can dramatically reduce storage footprint of the search index by 98% (from 3200 bits to 56 bits) with virtually no loss of accuracy. To the best of our knowledge no prior work has addressed the compact representation of descriptors for SBIR, with previous bag of visual words based SBIR system requiring between 3–10k bits per image.

Future work could explore alternative data augmentation strategies during training to further mitigate over-fitting on the relatively small free-hand sketch datasets currently available (e. g. adopting the stroke omission policy of Yi Li , QMUL) and alternative strategies for cleaning and decluttering the Internet sourced images for the triplets. Triplet ranking models perform optimally when a wide margin is present between positive and negative exemplars, and more sophisticated strategies for triplet proposal may hold value. A popular technique explored also in photo-based visual search has been the use of hard negative examples to further refine triplet performance (Gordo et al., 2016; Radenović et al., 2016). We have explored a similar approach, for our cross-domain problem; bootstraping the CNN using the embedding trained in a previous epoch to retrieve positive and negative images from the training set. So far this has met with limited success; possibly due to the cross-modality of the problem causing greater spread between anchor and positive/negatives in the embedding, or due to high levels of clutter in some edge maps. It appears that naïve application of hard negative mining is inappropriate for a sketch-photo regression, reducing overall model performance – or if using semi-hard negatives yielding negligible improvement to overall mAP. On that basis it would also be elegant to learn rather than prescribe the pre-processing (edge extraction) step, perhaps reducing clutter in doing so. We note that virtually all prior SBIR work adopts such an edge extraction step, and so this is likely a fruitful area for experimentation.

Acknowledgements

This research was supported in part by an EPSRC (grant #EP/N509772/1) doctoral training grant, and in part by FAPESP (grants #2016/16111-4 and #2015/26050-0). We gratefully acknowledge Nvidia Corp's donation of GPU hardware to support this work.

References

Arbelaez, P., Maire, M., Fowlkes, C., Malik, J., 2011. Contour detection and hierarchical image segmentation. IEEE Trans. Pattern. Anal. Mach. Intell. 33 (5), 898–916.
Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24 (4), 509–522. doi:10.1109/34.993558.
Bimbo, A.D., Pala, P., 1997. Visual image retrieval by elastic matching of user sketches. IEEE Trans. Pattern Anal. Mach. Intell. 19 (2), 121–132.
Bui, T., Collomosse, J., 2015. Scalable sketch-based image retrieval using color gradient features. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1–8.
Cao, Y., Wang, C., Zhang, L., Zhang, L., 2011. Edgel index for large-scale sketch-based image search. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 761–768.
Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 1. IEEE, pp. 539–546.
Cicso, 2014. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019. White Paper.
Collomosse, J.P., McNeill, G., Qian, Y., 2009. Storyboard sketches for content based video retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 245–252.
Collomosse, J.P., McNeill, G., Watts, L., 2008. Free-hand sketch grouping for video retrieval. Intl. Conf on Pattern Recognition (ICPR).
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 1, pp. 886–893.
Eitz, M., Hays, J., Alexa, M., 2012. How do humans sketch objects? ACM Trans. Graph. (Proc. SIGGRAPH) 31 (4), 44:1–44:10.
Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M., 2009. A descriptor for large scale image retrieval based on sketched feature lines. In: Proc. SBIM, pp. 29–36.
Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M., 2011. Sketch-based image retrieval: benchmark and bag-of-features descriptors. IEEE Trans. Vis. Comput. Graph. 17 (11), 1624–1636.

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., 1995. Query by image and video content: the QBIC system. Computer 28 (9), 23–32.
Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T., 2013. Devise: a deep visual-semantic embedding model. Neural Information Processing Systems (NIPS).
Gordo, A., Almazán, J., Revaud, J., Larlus, D., 2016. Deep image retrieval: learning global representations for image search. In: Proceedings of the European Conference on Computer Vision, pp. 241–257.
Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, 2. IEEE, pp. 1735–1742.
Hu, R., Barnard, M., Collomosse, J.P., 2010. Gradient field descriptor for sketch based retrieval and localization. In: Proc. ICIP, 10, pp. 1025–1028.
Hu, R., Collomosse, J., 2013. A performance evaluation of gradient field HOG descriptor for sketch based image retrieval. Comput. Vis. Image Understanding 117 (7), 790–806. doi:10.1016/j.cviu.2013.02.005.
Jacobs, C.E., Finkelstein, A., Salesin, D.H., 1995. Fast multiresolution image querying. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, pp. 277–286.
James, S., Collomosse, J., 2014. Interactive video asset retrieval using sketched queries. In: Proceedings of Conference on Visual Media Production (CVMP).
Jegou, H., Douze, M., Schmid, C., 2009. Packing bag-of-features. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, pp. 2357–2364.
Jegou, H., Douze, M., Schmid, C., 2011. Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell. 33 (1), 117–128.
Jegou, H., Douze, M., Schmid, C., Perez, P., 2010. Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3304–3311.
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM, pp. 675–678.
Kato, T., 1992. Database architecture for content-based image retrieval. In: SPIEIS&T 1992 Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics, pp. 112–123.
Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems.
Lam, L., Lee, S.-W., Suen, C.-Y., 1992. Thinning methodologies-a comprehensive survey. IEEE Trans. Pattern Anal. Mach. Intell. 14 (9), 879.
Lee, Y.J., Zitnick, C.L., Cohen, M.F., 2011. Shadowdraw: real-time user guidance for freehand drawing. ACM Trans. Graph. (TOG) 30 (4), 27:1–27:10.
Li, Y., Hospedales, T.M., Song, Y.-Z., Gong, S., 2015. Free-hand sketch recognition by multi-kernel feature learning. Comput. Vis. Image Understanding 137, 1–11. doi:10.1016/j.cviu.2015.02.003.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110. doi:10.1023/B:VISI.0000029664.99615.94.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. ICLR Workshop.
Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, 2. IEEE, pp. 2161–2168.
Perronnin, F., Liu, Y., Sanchez, J., Poirier, H., 2010. Large-scale image retrieval with compressed fisher vectors. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3384–3391.
Ponti, M., Nazaré, T.S., Thumé, G.S., 2016. Image quantization as a dimensionality reduction procedure in color and texture feature extraction. Neurocomputing 173, 385–396.
Qi, Y., Song, Y.-Z., Xiang, T., Zhang, H., Hospedales, T., Li, Y., Guo, J., 2015. Making better use of edges via perceptual grouping. In: IEEE Conference on Computer Vision and Pattern Recognition.
Qi, Y., Song, Y.-Z., Zhang, H., Liu, J., 2016. Sketch-based image retrieval via siamese convolutional neural network. In: Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, pp. 2460–2464.
Radenović, F., Tolias, G., Chum, O., 2016. CNN image retrieval learns from BoW: unsupervised fine-tuning with hard examples. In: Proceedings of the European Conference on Computer Vision, pp. 3–20.
Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S., 2014. CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806–813.
Rippel, O., Paluri, M., Dollar, P., Bourdev, L., 2016. Metric Learning with Adaptive Density Discrimination. stat 1050, 2.
Saavedra, J.M., Barrios, J.M., 2015. Sketch based image retrieval using learned keyshapes. In: Proceedings of the British Machine Vision Conference.
Saavedra, J.M., Bustos, B., 2014. Sketch-based image retrieval using keyshapes. Multimedia Tools Appl. 73 (3), 2033–2062.
Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823.
Sciascio, E.D., Mingolla, G., Mongiello, M., 1999. Content-based image retrieval over the web using query by sketch and relevance feedback. In: Visual Information and Information Systems. Springer, pp. 123–130.
Shechtman, E., Irani, M., 2007. Matching local self-similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. doi:10.1109/CVPR.2007.383198.
Sun, Y., Chen, Y., Wang, X., Tang, X., 2014. Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp. 1988–1996.
Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: a neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE.
Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y., 2014. Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393.
Wang, X., Gupta, A., 2015. Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.
Weinberger, K.Q., Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. 10, 207–244.
Wohlhart, P., Lepetit, V., 2015. Learning descriptors for object recognition and 3D pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3109–3118.
Yi Li (QMUL), S.G., Yi-Zhe, Song, 2013. Sketch recognition by ensemble matching of structured features. In: Proceedings of the British Machine Vision Conference. BMVA Press.
Yu, Q., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T.M., Loy, C.C., 2016. Sketch me that shoe. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE.
Yu, Q., Yang, Y., Song, Y.-Z., Xiang, T., Hospedales, T.M., 2015. Sketch -a-net that beats humans. In: Proceedings of the British Machine Vision Conference. IEEE.

Please cite this article as: T. Bui et al., Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.007