Diagram Image Retrieval using Sketch-Based Deep Learning and Transfer Learning

Manish Bhattarai[1,2,*], Diane Oyen[2], Juan Castorena[2], Liping Yang[1], and Brendt Wohlberg[2]

[1] University of New Mexico, Albuquerque, NM, USA
[2] Los Alamos National Laboratory, Los Alamos, NM, USA
[*] **Corresponding author**: Manish Bhattarai, ceodspspectrum@lanl.gov

Abstract

Resolution of the complex problem of image retrieval for diagram images has yet to be reached. Deep learning methods continue to excel in the fields of object detection and image classification applied to natural imagery. However, the application of such methodologies applied to binary imagery remains limited due to lack of crucial features such as textures,color and intensity information. This paper presents a deep learning based method for image-based search for binary patent images by taking advantage of existing large natural image repositories for image search and sketch-based methods (Sketches are not identical to diagrams, but they do share some characteristics; for example, both imagery types are gray scale (binary), composed of contours, and are lacking in texture). We begin by using deep learning to generate sketches from natural images for image retrieval and then train a second deep learning model on the sketches. We then use our small set of manually labeled patent diagram images via transfer learning to adapt the image search from sketches of natural images to diagrams. Our experiment results show the effectiveness of deep learning with transfer learning for detecting near-identical copies in patent images and querying similar images based on content.

1. Introduction and motivation

The patent industry involves the management and tracking of an enormous amount of data, much of which takes the form of scientific drawings, technical diagrams and hand sketched models. The comparison of figures across this dataset and subsequent retrieval based on similarity in real-time is extremely challenging [16], [30], [31]. We aim to track the spread of technical information by finding copies and modified copies of technical diagrams in patent databases and academic journals. Machine Learning (ML), and especially Deep Learning (DL) techniques offer the possibility of performing thousands of diagram/diagram comparisons across multiple databases in seconds. We present an ML approach that offers a high comparison accuracy with very little training data and, given a specific diagram and image of interest, under single shot and zero shot conditions can scan a database and retrieve all of the closest matches in that database for further review. Here, we present a deep learning approach that takes advantage of existing natural image repositories for image search and sketch-based methods applied to binary patent imagery.

The success of conventional CNN frameworks is widely acknowledged in image classification and cross-domain reconstruction applied to natural imagery when images have contextually rich information such as texture and pattern [3],[37]. These state-of-the-art frameworks fail when applied to diagrams, due to their contextually poor imagery. Standard One [29], Zero-Shot(ZS) [26] and Few-Shot(FS) [25] techniques, originally developed for small datasets struggle to perform well on diagram-type imagery due to the domain variation and huge non-overlap in representation across these domains. Most technical diagrams, sketches and scientific drawings found in patents are binary images. They lack significant features such as texture, color and contrast. Also, there is structural variation because of rigid body transformations such as translation, rotation or perspective variations (i.e. viewpoint change). Classical image processing and computer vision tools such as key-point matching do not perform well given such transformations [12]. Typically, DL performance is robust against such rigid body transformations when trained with data augmentation techniques [22]. However, due to the lack of sufficiently labeled patent image data available, DL models can easily over-fit when trained on the small datasets typical of patent-related imagery.

Domain generalization[17] and domain adaptation [19] techniques are gaining popularity as methods to address this

data gap. Domain generalization provides a method to generalize the trained model over a broader dataset. Here, we apply the concept of domain generalization by pre-training an unsupervised DL model on a large set of sketches generated from natural images. We aim to achieve a generalized representation of the latent space with the edge maps and then project the target patent dataset to this domain-invariant representation where differences between training domains are minimized by incorporating the proper loss functions. We explore this method in both few-shot and zero-shot conditions where the model is able to generalize the matches and make similarity predictions based on a small subset of the dataset for training. The model learns to recognize unseen matching pairs based on knowledge acquired from training of labeled similarity pairs.

While most image retrieval methods and algorithms are designed around natural imagery, sketch-based retrieval [13] provides promise as a means to further image retrieval related to patents. Our methods further extend their approach through the following steps:

1. We use deep learning to generate sketches from natural images (using existing natural image repositories for image retrieval/image search/ image comparison).

2. The large dataset of sketches created in (1) is used in the training for image retrieval (because if the original natural images match, we assume the corresponding sketches will match as well).

3. The unsupervised deep learning model is trained on the sketches dataset for domain generalization.

4. We use transfer learning (our small labeled subset of data for image query is used at this stage) to complete the image retrieval task based on the model trained in (3).

We show that even under zero-shot and one-shot conditions, this framework surpasses classical retrieval frameworks for retrieval of similar binary images.

2. Related work

The requirements of a patent image retrieval system include full-image, sub-image, category-based image, rotation-, scale- and affine-invariant image searches, real-time performance, scalability, on-line learning, and semantic level interpretation. Although the combined set of requirements present significant challenges, we aim to address most of them in our approach.

Content-based Image Retrieval (CBIR) makes use of low level visual features such as color, edges, texture, and shape to represent and retrieve images [1, 24, 37]. **Relational skeletons** [8], consider features such as relational angle and relational position between lines. The use of line segments in representing the image makes this approach sensitive to rigid body transformations such as rotations, translations and scaling. The **Edge Orientation AutoCorrelogram (EOAC)** approach used in the US patent retrieval system PATSEEK [28] claims to be insensitive to translation and scaling; however the approach is computationally expensive and the complexity grows with the feature vector size. The use of user-defined thresholds makes the approach scale variant. The **Contour Description Matrix (CDM)** approach [36] uses canny edge detection for extracting contour information followed by converting each edge point to a polar coordinate system. While this approach is invariant to rigid body transformations, the size of the CDM is dependent on image resolution and the resulting processes are inefficient both computationally and memory-wise. The **Adaptive Hierarchical Density Histogram (AHDH)** method [23] along with the retrieval framework PATMEDIA [31] exploits both local and global content. It uses both content-based (i.e image-based) as well as concept-based (text-based) retrieval and claims joint retrieval using both text and image give better retrieval performance. The algorithm calculates the adaptive hierarchical density histogram by computing the density of black pixels on a white plane after reducing noise and normalizing at the pre-processing stage. The ADHD process is made to retrieve the images belonging to the same category in the database and fails to retrieve similar images belonging to a different category. Besides, one needs to also manually set two different thresholds to make the system scale invariant which contradicts the idea of scale invariance. **Fisher vectors** based patent retrieval [2] uses Fisher vectors [20] to represent patent images as low level features. For a pair of images, a dot product of fisher vectors is computed to measure the similarity between them. Similar to the ADHD approach, this approach does categorical based retrieval instead of similarity based retrieval.

3. Datasets

The dataset used to train and test our model is taken from a patent image search benchmark [30]. About 2000 sketch-type images are manually extracted from approximately 300 patents belonging to A43B and A63C IPC subclasses and contain types of foot-wear or portions thereof (henceforth termed ”concepts”). The dataset consists of 8 concepts for this domain: cleat, ski boot, high heel, lacing closure, heel with spring, tongue, toe cap and roller blade. The details for the dataset can be found in [30]. The concepts dataset contains many dissimilarities within each class and is not suitable to train a classifier model to be used as a retrieval and matching framework.To ground-truth this concept dataset, we evaluated image similarity through manual pairwise comparisons made by three different non-experts

and then determined a median out of all similarities. The pairwise similarity was quantified in the score range of 0-5 where, 5 - Same match, 4 - Slightly different, 3 - different perspective, 2 - sub-image, 1 - slightly different sub-image and 0 - dissimilar.

We used the UT Zappos50K shoe dataset [35] and the Generative Fashion dataset [21] to generate the sketches for domain generalization. The first dataset contains a total of 50K catalog images that were collected by Zappos.com while the second contains 293K high resolution fashion images. Using these two datasets, a retrieval performance was measured on the concepts dataset and fashion-MNIST [33] dataset respectively. The Fashion-MNIST dataset contains 70k images of gray scale fashion products in 10 categories.

4. Methods

Labeled benchmark datasets of natural images are easily accessible online, but labeled datasets of patent diagrams are more limited. To generate the sketches/edge-maps intended for usage as our custom shoe training dataset, we process the collection of natural images through the use of the Holistically Nested Convolutional Neural Nets(HCNN) [34]. We train a Variational Auto-Encoder (VAE) [9], an unsupervised representation learning model, to approximate the distribution of the newly generated sketch dataset with a multi-dimensional Gaussian distribution with finite mean and variance. Once the model learns the representation of the data, we reuse this model via transfer learning on our small dataset for domain generalization [18]. The idea of domain generalization is to learn from one or multiple training domains, to extract a domain-agnostic model which can be applied to an unseen domain. We show that, on passing the dataset through this learned model, it is able to achieve a minimal clustering of similar matches of the dataset. We augment this model with extra blocks of neural nets to construct a Siamese framework [10] for fine tuning of the features on the latent space using triplet loss [5] that bring likely samples closer and push dissimilar samples farther away. During the training of the Siamese framework, the augmented block is fine tuned with a small subset of the similarity matrix from the entire dataset. At the test phase, the samples that are used to query may/may not have been present during the training. If the similarity metric corresponding to the queried sample was used during training, then it is called Few-Shot/One-shot learning whereas if the no similarity metric corresponding to the queried sample was used during training, then it is called Zero-Shot learning. This applicability of one shot and zero shot retrieval with our framework relies on the knowledge gained during the domain generalization followed by intelligent fine tuning of the features.

Once we have ideal clustering of the samples in the latent space via domain generalization and Siamese triplet loss

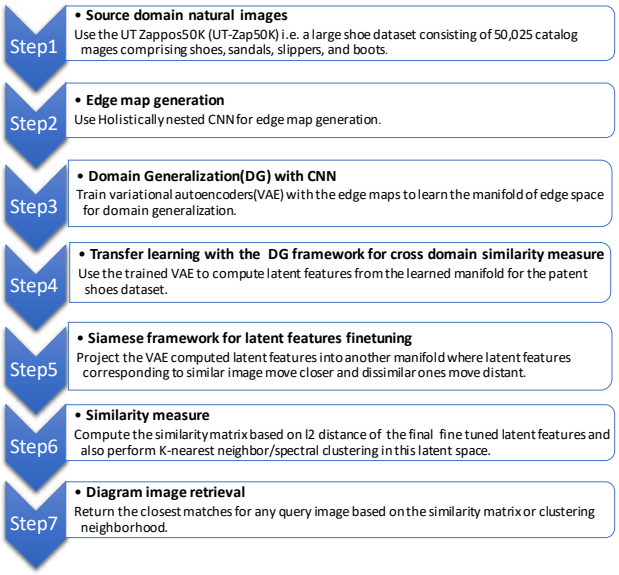


Figure 1: Proposed model flow-chart

based fine-tuning, we can use k-nearest neighbour(kNN) clustering to return the more closely matched pairs. This could be incorporated into a retrieval tool to return the best set of matching images from the queried database.

To measure image similarity between two images X, Y with corresponding pixels $\{x\} \in X$ and $\{y\} \in Y$ we use the mathematical expression:

$$S(X,Y) = \sum_{x \in X} \sum_{y \in Y} K(x,y) \tag{1}$$

$$= \sum_{x \in X} \sum_{y \in Y} \phi(x)^T \phi(y) \tag{2}$$

$$= \psi(X)^T \psi(Y) \tag{3}$$

where $K : \mathbb{R} \rightarrow \mathbb{R}$ is the operator denoting pixel similarity. Note that Eq. 1 is equivalent to the Kernel factorization of [6] where image similarity is computed from features defined by the operator $\phi : \mathbb{R} \rightarrow \mathbb{R}$. Alternately, the similarity is a dot product between the transformed features as described by the function $\psi()$. These three expressions (1), (2) and (3) dictate the process of feature extraction, feature encoding and aggregation and database indexing respectively.

4.1. Holistically Nested Convolutional Neural Nets (HCN) for edge map generation

We exploit a state-of-the-art edge detection algorithm called Holistically Nested Convolutional Neural Nets (HCN) [34] to generate the edge maps. This model utilizes an end-to-end deep CNN framework for image to image prediction where the input and output are natural image and edge map respectively.

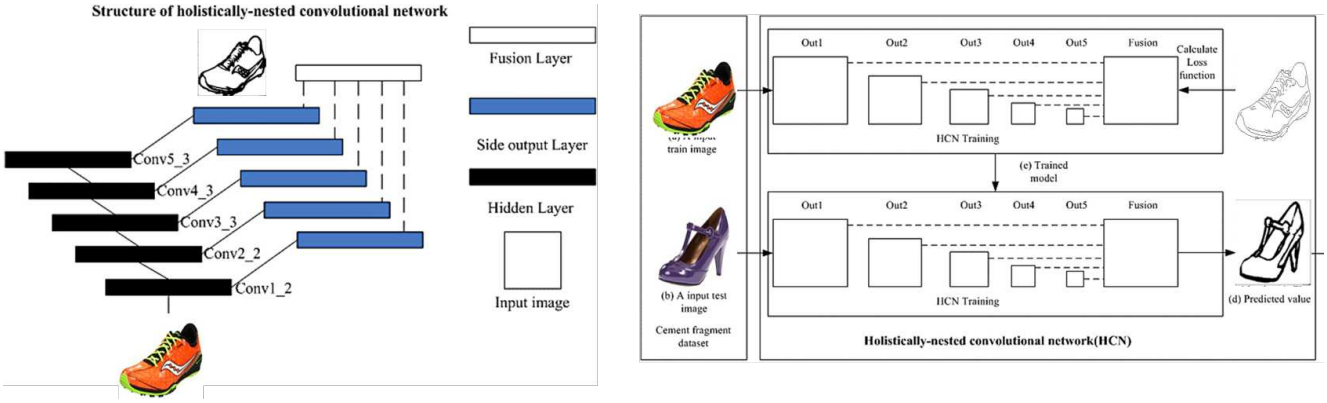


Figure 2: An overview of the Holistically Nested Edge Detection framework

The model comprises a modified VGG16 network [11] where the final pooling and fully connected layer is pruned. A deep supervision is established by connecting the side output layer to the last convolutional layer in each stage, Conv1_2, Conv2_2, Conv3_3, Conv4_3, and Conv5_3, respectively. A convolutional layer with a kernel size of 1 is operated on the output of each of the previous layer outputs to compute side outputs which are all then connected to a final fusion layer. The framework is trained with image sketch pairs and then tested with the shoes natural images. Figure 2 demonstrates the HCNN framework for edge map generation. The overall loss function is given by

$$L(I, G, W, w) = L_{side}(I, G, W, w) + L_{fuse}(I, G, W, w) \tag{4}$$

Where,
L_{fuse} = fusion layer loss function,
L_{side} = side output layer loss function,
I = raw input image,
G = ground truth binary segmentation map,
W = collection of all other network layer parameters,
$w = w(1), w(2), \ldots\ldots.w(M)$: corresponding weights for each side output layer

4.2. Domain generalization via Variational Auto Encoder(VAE)

We adopt the concept of domain generalization for representation learning. This refers to the learning representation of a domain dataset which makes it easier to extract significant information when building the matching framework. This kind of representation learning is usually done in unsupervised settings by leveraging the potential of the excess unlabeled dataset. Domain generalization tries to approximate the latent space/manifold over which the data of interest can be projected for categorization,clustering, or matching.

We aim to combine a self supervised data representation achieved via a pre-training process and fine tune the model through transfer learning. We use a Variational Auto-Encoder (VAE) which is implemented on an explicit reconstruction loop that focuses on achieving per-pixel reconstruction. This VAE is trained for the purpose of unsupervised data representation and uses the encoder framework in a Siamese framework to achieve a benchmark performance of image matching/retrieval via transfer learning.

The VAE builds generative models of complex distributions of the sketched shoes dataset. It uses a CNN based function approximator to approximate an otherwise intractable function. The encoder encodes the input data into the mean and variance statistics of the latent space and then samples data points from the Gaussian distribution computed from the statistics. The decoder tries to reconstruct the input data based on the sampled points. The framework trains in an end-to-end fashion where the objective for the encoder is to generate the statistical encoding in such a way that the difference between the input image and the reconstructed image are minimized.

The VAE is incorporated to generate an observation x from some hidden variable z such that $p(z|x)$(intractable distribution) is approximated by another distribution $q(z|x)$ via approximate inference. With the objective to minimize the KL divergence between these two distributions $q(z|x)$ and $p(z|x)$ and also minimize the reconstruction error, we can write the overall loss function as

$$\theta(x) = KL(q_\phi(z|x)||p(z|x)) + L(p_\theta, q_\theta), \tag{5}$$

where

$$L(p_\theta, q_\theta) = E_{q_\phi(z|x)}[logp_\theta(x, z) - logq_\theta(z|x)] \tag{6}$$

The first term represents the reconstruction error reconstruc-

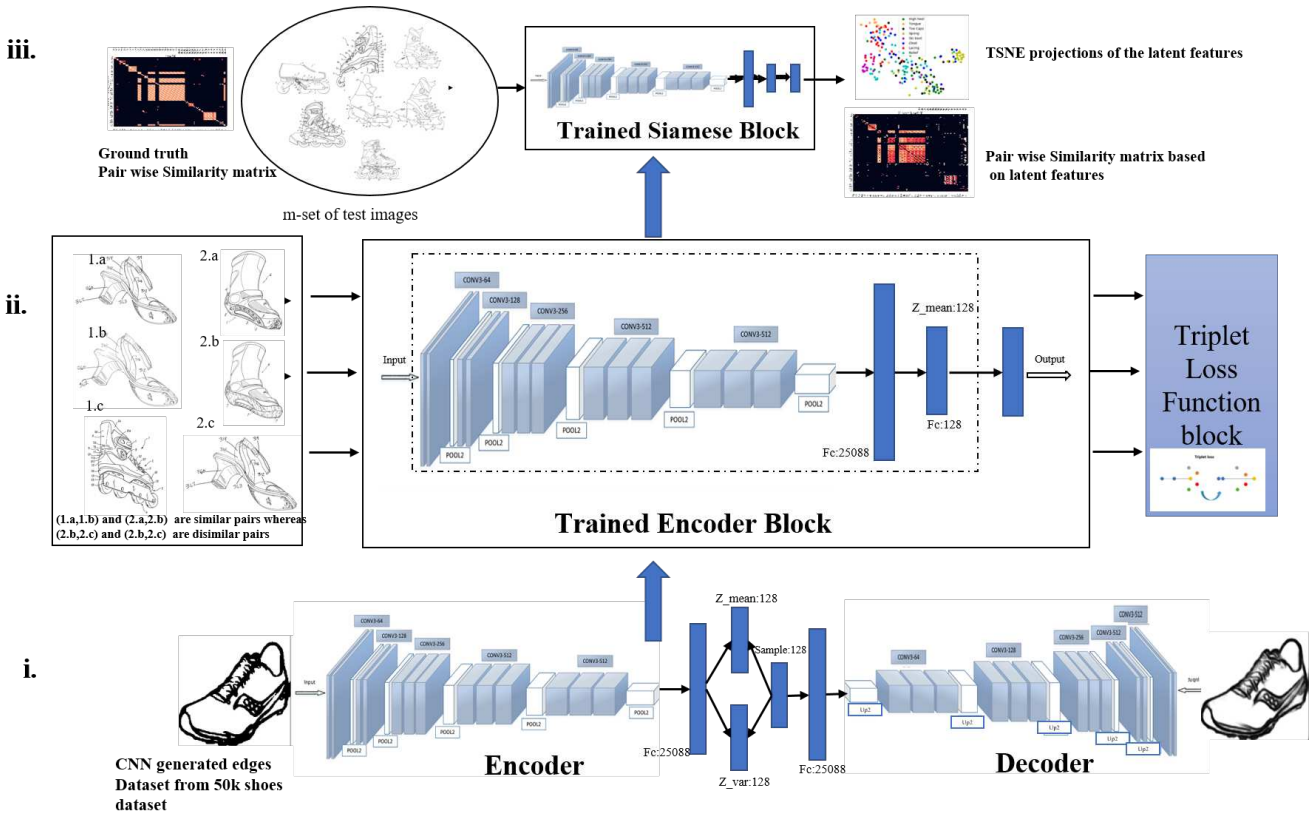


Figure 3: An overview to the single-shot/zero-shot framework. **i.** Learning the latent representation of the 50K edge dataset with the Variational Auto-Encoder(VAE) for domain generalization. **ii.** Transfer learning of the trained Encoder block into the Siamese framework for training with concept dataset. **iii.** t-distributed stochastic neighbor embedding(t-SNE) projection and similarity matrix computation based on latent features generated by the trained Siamese block.

tion likelihood and the second term ensures that our learned distribution q is similar to the true prior distribution p.

4.3. Single shot/zero shot training for image retrieval using a Siamese framework

Once we have obtained the optimal latent representation of the latent space with the VAE, the encoder framework can be used to extract the optimal latent representation for our small dataset. To achieve a better cluster between similar image pairs, we implement triplet loss for a Siamese network. To compute triplet loss, we consider an anchor i.e. a reference from which the distance will be calculated to a positive sample (e.g. sample with a large paired similarity score) and negative sample (e.g. sample with 0 paired similarity score). If we consider the anchor, positive and negative sample images as x_i^a, x_i^p, x_i^n and corresponding embedding vectors as f_i^a, f_i^p, f_i^n, then the triplet loss $l_{triplet}$ is given as

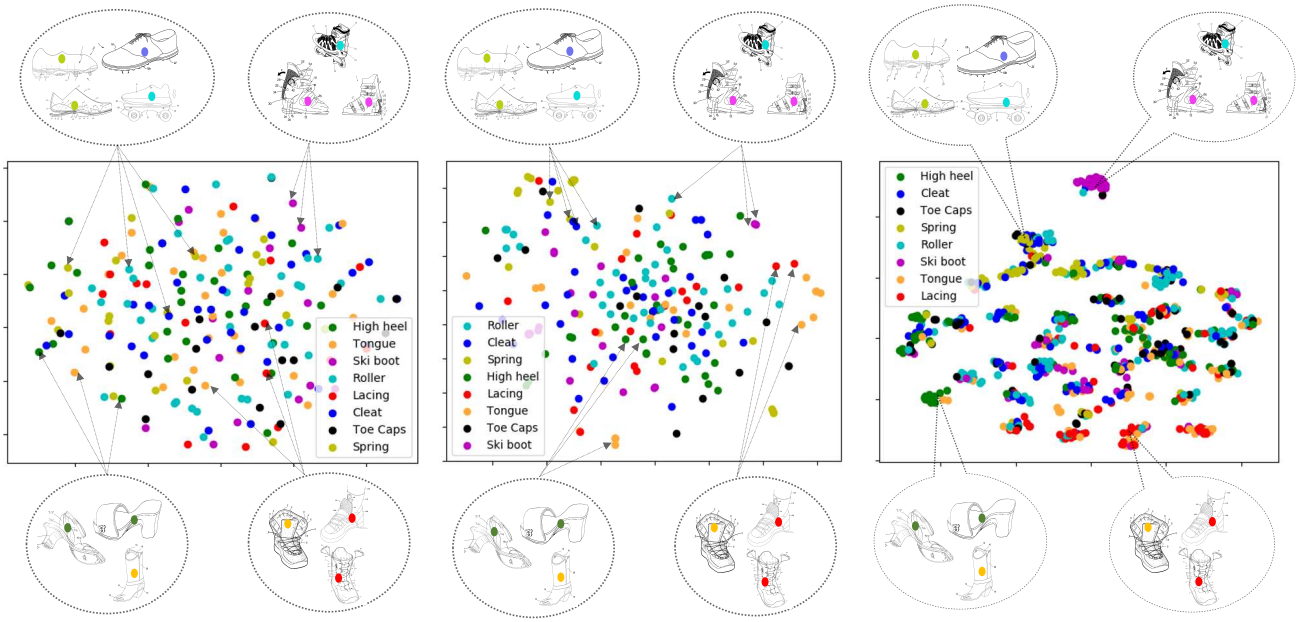
$$l_{triplet} = 1/N \sum_{i=0}^{N} max(0, ||f_i^a - f_i^p||_2^2 - ||f_i^a - f_i^n||_2^2 + \alpha) \tag{7}$$

Where N is the batch size and α is a constant factor.

Figure 3 gives a broader overview of the implemented methodology. In figure 4, we can see TSNE embedding corresponding to latent features at different stages of the framework.

5. Experiments and results

We used the Keras framework with a Tensorflow backend to train and fine tune the proposed model. All experiments were conducted on the Darwin cluster at the Los Alamos National Labs. This cluster is equipped with Intel(R) Xeon(R) Gold 6138 CPUs and 8 GeForce RTX 2080Ti GPUs. The dataset for training was constructed from the similarity matrix as triplets. The split was done 60%(training+validation) and 40%(test). A relatively larger test set was chosen to measure the performance of the model



(a) (b) (c)

Figure 4: (a)TSNE Projection of image space (b)TSNE representation of VAE output feature space (c)TNSE representation of Siamese tuned output feature space

under One-Shot and Zero-Shot conditions.

The implementation of the HCNN framework was borrowed from [1]. Considering the pertained model, the model used to generate the pseudo-sketches from the natural image datasets as discussed in the Dataset section.

We constructed the deep VAE from the standard VGG16 architecture by trimming the fully connected layer of VGG16 and augmenting the left model with a single layer VAE model to form an encoder and then combining with an equivalent decoder as shown in bottom of figure 3. We validated the VAE model with the following hyper-parameters: a number of convolutional layers and encoding dimensions for mean and variance. The configuration including a 5-layered Convolutional block and 128 encoding dimensions achieved the best reconstruction accuracy on the pseudo sketch datasets. We used the batch size of 64, Adam optimizer and an initial learning rate of 0.001 for training the VAE.

Next, the trained encoder model from the previous step was augmented with a fully connected block(FC) to construct a Siamese network whose triplet loss is computed from the set of three input images. We experimented with additional loss functions including the Contrastive loss [4], Constellation loss[15] and n-pair loss[27]. Triplet loss provided the best similarity performance. We also investigated the performance of the Siamese framework by i) training the FC while freezing the encoder block and ii) training both FC and encoder in end-to-end fashion. We observed that the second approach achieves superior performance. This is due to a larger allowance for parameter learning for tuning the feature space. From figure 4, considering different groupings of similar items, the distance between sample points decrease in the feature space from a to c as shown by the t-Distributed Stochastic Neighbor Embedding(TSNE)[14] projection. The data points are randomly distributed in the original pixel space and the pre-trained VAE achieved an improved level of closeness between similar samples without any knowledge of the patent dataset. Furthermore, with fine tuning utilizing the Siamese framework, the grouping of the samples based on their similarity is achieved.

Before any retrieval task, we construct a full similarity matrix that encodes all computed pairwise similarities between the elements of the dataset. Pairwise similarities can be quantified as the cosine similarity or as the Euclidean distance between the features collected at the output of the Siamese layer followed by normalization of values to fall on a scale of 0-5. For the patent shoes concept dataset, the similarity matrix is of the size 1042 X 1042. To query any image from the row of the matrix, we process the corresponding columns and sort them based on the score (i.e highest to lowest score) and return the first k elements of the sorted array. Figure 5 is an example comparing ground truth and

[1] https://github.com/moabitcoin/holy-edge

Structural Similarity Index Measure(SSIM) [7] based similarity matrices for the retrieval of a queried image. However, if one needs to perform retrieval for additional images outside of the database, then the results of the query will be based on a newly computed similarity matrix that includes the newly added images.Figure 6 demonstrates the retrieval results under One-Shot and Zero-Shot conditions. To generalize our retrieval framework to a dataset with no

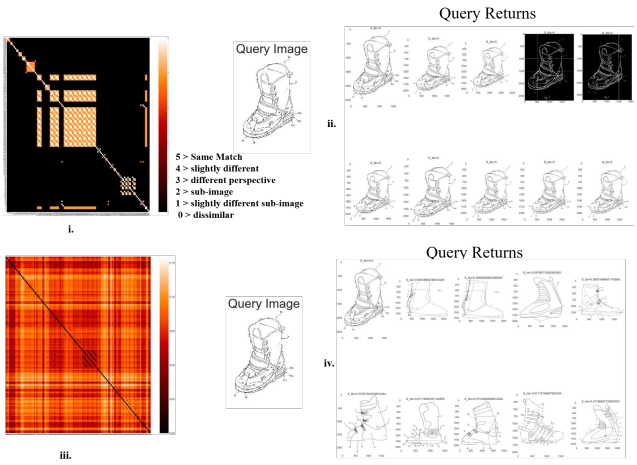


Figure 5: i) The pairwise ground truth similarity matrix corresponding to the class ski boot. ii) Illustration of the database returns for the given query image. iii) Pairwise Similarity Matrix based on the Structural Similarity Index Measure (SSIM) between the images iv) Query returns based on the similarity matrix iii.

available baseline similarities, we train using a binary similarity matrix with a score of 0 for intra-class and a score of 1 for inter-class images. This framework is likewise trained in zero-shot and one-shot conditions. The output features of the Siamese network in the test dataset were clustered with k-NN to obtain the nearest k features given a test input. To measure retrieval performance we use the mean average precision (MAP) score computed as:

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{Ave}(P(q))}{Q} \tag{8}$$

over all retrievals. Here, Q is the number of queries and $\text{Ave}(P(q))$ is the average of the precision score for each query q.

To justify the effectiveness of the proposed approach, we first performed retrieval on variants of our model and measured the MAP results on the concept patent dataset. When the VAE was first trained on natural images instead of sketches followed by the training of the Siamese framework on top of that model, we achieved an overall MAP of 0.75 for the first 10 retrievals. We also tried training a randomly initialized encoder block for the Siamese network and the MAP dropped to 0.6 for the same retrieval set. Implementing our proposed approach, the overall MAP for the same retrieval set resulted in a MAP score of 0.83.

For baseline comparisons, we use the standard SSIM and Goldberg(GB) similarity [32] to compute the structural similarity between all the image pairs on the test data partition and then return the set of k-NN images with the highest similarity measure. Table 1 summarizes the retrieval performance of the proposed framework in comparison to SSIM method for two benchmark datasets: 1) the concept shoe and 2) fashion-MNIST datasets. While SSIM is still used for image matching, classification and retrieval, GB is used as a tool for elastic search in larger datasets[2][3]. Here, retrieval scores are measured with MAP estimates for 10, 20 and 30 retrieved images per image query. Figure 7 instead shows a more detailed trend of the retrieval performance as a function of the number of retrieved images where our proposed framework outperforms SSIM and GB based retrieval in both datasets. Notice also that our method performance decreases smoothly with increases in items retrieved irrespective of image view-point and intensity variations.

In both Zero-Shot and One-Shot retrieval cases in the concept dataset, performance of the proposed approach is significantly better than SSIM and GB which drops in performance exponentially with increase in the number of retrieved items. Also, note that performance is lower in the Zero-Shot compared to the One-Shot framework caused mainly because of the additional knowledge acquired by the model with regard to the query set in the One-Shot case. Also, note that for fashion-MNIST similar retrieval performances are achieved in both proposed and SSIM methods. This is mainly due to high intra-class similarity, the lack of multiple viewpoints and the binary scoring used in this case. In contrast, for the concept patent dataset, the similarity scoring which ranges from 0-5 complicates the retrieval process for any other models that are not trained with such a scoring scheme. In both datasets, GB fails to perform well for the retrieval in binary and gray-scale images in contrast to its efficacy for retrieval in large RGB datasets.

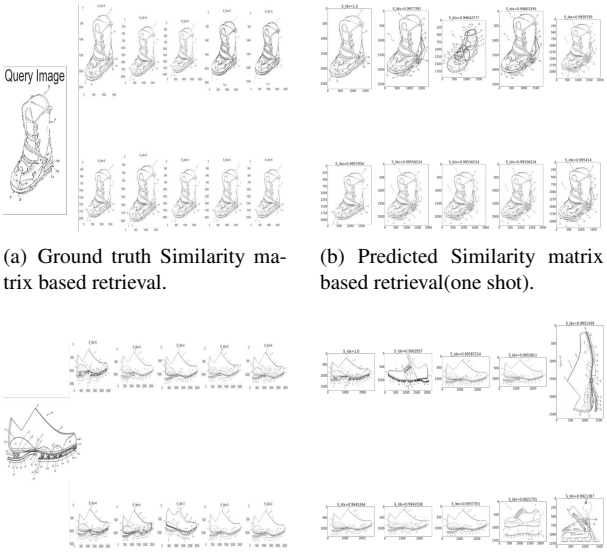
6. Conclusion and future work

We have demonstrated that domain generalization with the edge maps-based inductive short term learning and latent space fine-tuning based transductive long term learning aids to improve retrieval performance. This two-step process helps to fine tune the feature space by appropriately learning the data manifold. This provides a more meaningful structure of technical diagrams which naive image processing/computer vision techniques are unable to extract. Also, we have demonstrated image retrieval using the do-

[2] https://github.com/EdjoLabs/image-match
[3] https://github.com/dsys/match

Table 1: Retrieval Performances

Dataset	mAP@10	mAP@20	map@30
Concept	**0.816(ZS on Proposed)**	**0.667(ZS on Proposed)**	**0.581(ZS on Proposed)**
	0.842(OS on proposed)	**0.721(OS on Proposed)**	**0.643(OS on Proposed)**
	0.348(GB)	0.280(GB)	0.247(GB)
	0.273(SSIM)	0.216(SSIM)	0.193(SSIM)
Fashion-MNIST	**0.86(Proposed)**	**0.81(Proposed)**	**0.77(Proposed)**
	0.807(SSIM)	0.770(SSIM)	0.750(SSIM)
	0.702(GB)	0.650(GB)	0.625(GB)



(a) Ground truth Similarity matrix based retrieval.

(b) Predicted Similarity matrix based retrieval(one shot).

(c) Ground truth Similarity matrix based retrieval.

(d) Predicted Similarity matrix based retrieval (zero shot).

Figure 6: Predicted Similarity matrix based retrieval

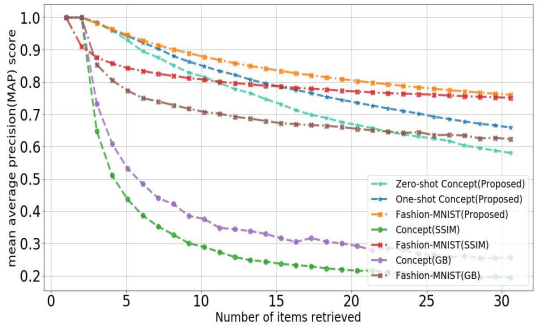


Figure 7: mean average precision(mAP) analysis on Concept Patent dataset(One-Shot and Zero-Shot) and Fashion dataset

main generalization concept on shoe patent images in One-Shot and Zero-Shot settings. We can extend this framework to other scientific drawings and patent images by pre-training the framework with related datasets.

To construct the similarity matrix to perform retrieval, we first computed the Euclidean distance/ Cosine similarity between the learning based deep features. These deep features were obtained by transfer learning of a deep learning model for patent images. It was observed that the learned deep features from the supervised classification based task or unsupervised latent representation did not reflect well on the retrieval performance as the learned features were more biased towards the learning objective the framework was trained for. In our approach, we build a pipeline where we fine tune the features obtained with transfer learning with the objective of achieving an improved similarity measure between the features corresponding to different images. This resulted in a better retrieval performance. Also, because of the difficulty involved in quantifying the retrieval performance on training the deep learning model, we instead use a similarity metric implemented via a scoring measure of similarity between image pairs to train the framework. In the future, we plan to implement a Bidirectional Generative Adversarial Networks (BiGANs) to learn generative models mapping from simple latent distributions to arbitrarily complex data distributions. This framework would be able to perform domain generalization across more broad image domains including natural images, sketches, scientific drawings and patent images.

7. Acknowledgments

Research supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory (LANL) under project number 20200041ER. MB supported by the LANL Applied Machine Learning Summer Research Fellowship(2019).

References

[1] S. Antani, R. Kasturi, and R. Jain. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern recognition*, 35(4):945–965, 2002. 2

[2] G. Csurka, J.-M. Renders, and G. Jacquet. Xrce's participation at patent image classification and image-based patent retrieval tasks of the clef-ip 2011. In *CLEF (Notebook Papers/Labs/Workshop)*, volume 2, 2011. 2

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 6

[5] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 3

[6] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008. 3

[7] A. Hore and D. Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010. 7

[8] B. Huet, N. J. Kern, G. Guarascio, and B. Merialdo. Relational skeletons for retrieval in patent drawings. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 2, pages 737–740. IEEE, 2001. 2

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[10] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015. 3

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 4

[12] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015. 1

[13] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2871, 2017. 2

[14] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 6

[15] A. Medela and A. Picon. Constellation loss: Improving the efficiency of deep metric learning loss functions for optimal embedding. *arXiv preprint arXiv:1905.10675*, 2019. 6

[16] M. Mogharrebi, M. C. Ang, A. S. Prabuwono, A. Aghamohammadi, and K. W. Ng. Retrieval system for patent images. *Procedia Technology*, 11:912–918, 2013. 1

[17] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013. 1

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1717–1724, 2014. 3

[19] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010. 1

[20] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[21] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 3

[22] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. 1

[23] P. Sidiropoulos, S. Vrochidis, and I. Kompatsiaris. Content-based binary image retrieval using the adaptive hierarchical density histogram. *Pattern Recognition*, 44(4):739–750, 2011. 2

[24] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1349–1380, 2000. 2

[25] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1

[26] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 1

[27] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016. 6

[28] A. Tiwari and V. Bansal. Patseek: Content based image retrieval system for patent database. In *ICEB*, pages 1167–1171, 2004. 2

[29] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 1

[30] S. Vrochidis, A. Moumtzidou, and I. Kompatsiaris. Concept-based patent image retrieval. *World Patent Information*, 34(4):292–303, 2012. 1, 2

[31] S. Vrochidis, A. Moumtzidou, G. Ypma, and I. Kompatsiaris. Patmedia: augmenting patent search with content-based image retrieval. In *Information Retrieval Facility Conference*, pages 109–112. Springer, 2012. 1, 2

[32] H. C. Wong, M. Bern, and D. Goldberg. An image signature for any kind of image. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I. IEEE, 2002. 7

[33] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 3

[34] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 3

[35] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014. 3

[36] Z. Zhiyuan, Z. Juan, and X. Bin. An outward-appearance patent-image retrieval approach based on the contour-description matrix. In *2007 Japan-China Joint Workshop on Frontier of Computer Science and Technology (FCST 2007)*, pages 86–89. IEEE, 2007. 2

[37] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 8(6):536–544, 2003. 1, 2