Dual Stage Generic Framework for Cross-Domain Content Matching and Retrieval

Anirudh Kannan[1], Dr. B Sathish Babu[1]

[1]*Department of Computer Science and Engineering, R V College of Engineering, Bengaluru, India*

---***---

Abstract - *Cross-Domain content refers to data from different source distributions. The task of matching two data points from two richly different distributions is a complex one, since traditional machine learning algorithms do not perform well at modelling a mapping function if the source and target domains are dissimilar. Although some Deep Learning models can help bridge this domain gap, they are computationally expensive and must be trained for long periods of time. In this paper we propose a robust Dual Stage Framework that can be used to perform this Cross-Domain mapping, requiring lesser computational power than existing methods. The first phase embeds the data points in dissimilar domains into a common latent space and the second phase maps the embeddings in the latent space with each other. The two phases can be trained independently, and hence require a lesser amount of computational power to produce robust results. The proposed hypothesis is verified on the sketchy database, producing strong results comparable with existing baselines. The proposed model demonstrates effective generalization and performance with limited resources.)*

Key Words: Cross-Domain, source distributions, domain gap, Dual Stage Framework, heterogeneous distributions, latent space, embeddings

1.INTRODUCTION

Human beings have an inherent ability to draw mapping and comparisons between dissimilar objects. This unique ability of humans to generalize existing knowledge with novel data allows us to adapt and draw conclusions quickly. Cross domain relations are naturally recognized by humans. However, this is not the case with Deep Learning Systems [1].

The boom of deep learning has led to an increased demand for data. However, the patterns and data encountered during the deployment of these models may not belong to the same domain that the training data belongs to. This is known as the covariate domain shift. It leads to poor translation of performance from training to testing. The ability of a model to generalize is also determined by its ability to handle covariate shifts.

When sets of data belong to different sources and have a domain gap, we call this a cross domain problem. Cross Domain data refers to data from different source distributions. Matching data from different domains is a complex problem since deep neural networks have poor generalizability to covariate domain shift.

To tackle the problem of Cross Domain Learning, several deep learning models have been developed including autoencoders and Siamese neural networks. One disadvantage of using these models is that they require a lot of data and computational power to be trained. For example, training a Siamese Network with dissimilar image pairs requires end-to-end training of a large network on multiple pairs of images. Generally, if there are N examples of each image, then there are O(N2) number of image pairs forming the training data. Siamese Networks also have several layers and hence training this model on the entire dataset requires a lot of time and computing resources (GPU's).

In this paper we propose a novel Dual Stage Framework for Cross Domain Content Matching which can generalize well to any content modalities. The first stage involves extraction of the latent space features from the input dataset. This is done in such a way that the features extracted from both distributions are of the same dimensionality. The second stage involves converging latent vectors which are similar (semantically) and diverging those which are dissimilar. It is important to note that the two deep neural networks of stage 1 and the network of stage 2 are trained independently. This reduces the computational resources required for training, since after the first stage, the dimensionality is reduced, leading to a simpler and smaller model being used for the second stage. Additionally, making cross domain pairs for training becomes simpler with lower dimensional features being used in place of the content directly. Although the number of pairs remains the same, the dimension of the data is reduced significantly, hence reducing consumption of GPU resources.

We demonstrate the efficacy of this framework on the Sketch Based Image Retrieval task [2]. The task of sketch-based image retrieval is to draw relevant images from a set of images given a hand drawn sketch. Here single channel sketches and multi-channel (RGB) images form the two domains. Since these are dissimilar distributions, cross domain matching [3] is required to additionally demonstrate the robustness of this approach with respect to performance on the baselines in terms of a metric known as mean average precision.

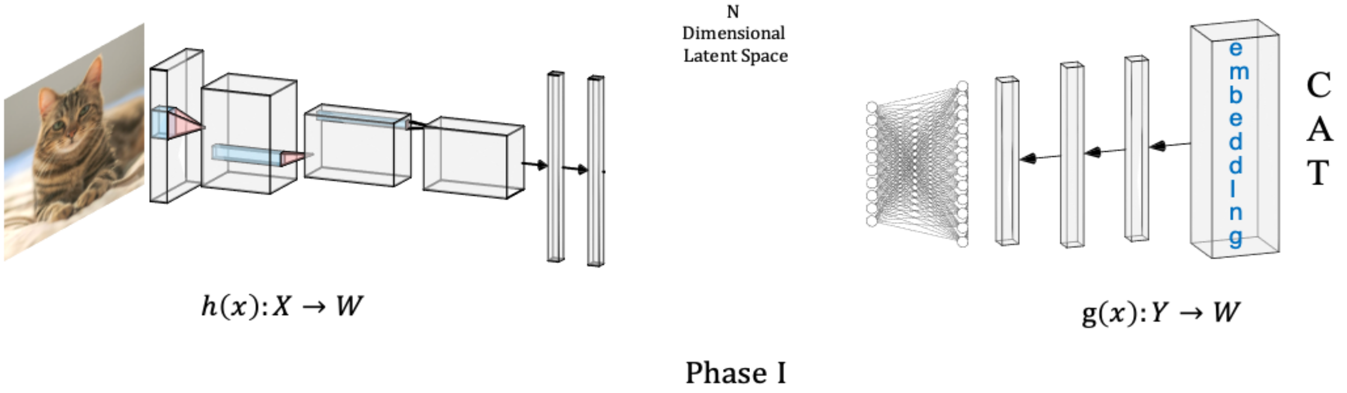


Fig -1: Stage 1 of the Framework

2. LITERATURE SURVEY

Prior work explored the use of Sketch Based Image Retrieval systems in trademark search and verification [4], clip art generation [5, 6] and in documents [7]. However, these works assumed that the sketches were a set of strokes and the geometric relation between them are considered for retrieval and matching [8, 9]. Clearly these techniques were very susceptible to perspectives, disturbances and occlusion.

Edgemap based approaches, including the use of the Canny edge detector and HOG [10] were also explored, however these failed to bridge the cross domain gap effectively. These techniques were used to convert both source distributions into a single type, however this approach was not effective due to loss of data and [11, 12] explored the use of Siamese Networks and deep learning for feature extraction. These techniques however train the entire network at once and require a lot of computational power. Our framework builds on these techniques, by proposing an architecture to reduce computational complexity and training time.

Current literature on Cross Domain Comparisons explores different approaches including disentangling domain features [13] and structure preserving learning [14]. While these techniques have been demonstrated to be effective, there is a requirement for training complex models. No current works explore training in conservative environments for robust performance, which is an important component in our framework.

Using the existing Siamese neural network for learning the task to perform cross domain translation, we modify the architecture in accordance with our proposed framework, including the separation of feature extractors as separate independent units as well as dimensionality reduction of the dataset and reduction of complexity of the Siamese Network.

3. FRAMEWORK

A detailed description of the two independent phases is given in the following sections. We assume that a data source is a set of input vectors belonging to a unique distribution.

3.1 Stage 1

The first phase of the Dual Phased Deep Learning Framework model involves training two mapping functions, each transforming the source data vectors into latent space embeddings. Consider two sources of data X1and X2. These data distributions are dissimilar and there is a significant domain gap between them. Consider two data points x1 and x2 belonging to distributions X1in m dimensional space and X2 in n dimensional space. Let W be the common latent space in which we aim to embed the data points.

The latent space W consists of all the vectors spanned by the feature embeddings of the data points of X1andX2.

Then we define two functions h(x1): X1->W and g(x2): X2->W such that

$$h(x) = w1 \text{ and, } \quad g(y) = w2,$$

where w1 and w2 are the output vectors from this mapping. The mapping functions are trained on the pairs of data such that if x1 and x2 vectors are close to each other as measured by Euclidean distance or cosine similarity, then their output vectors are also close to each other. The vice versa is also applicable.

Essentially, these mapping functions embed the data in such a way that clusters of semantically similar data points are formed. Source vectors, corresponding to the latent vectors within a cluster, are inherently similar to each other.

From an implementation perspective, this can be done by training a deep neural network to perform classification on the data, and by using the penultimate layer of the network, we get embeddings in an N dimensional space, where N is the number of neurons in the penultimate layer.

Another way to implement this, is by using an unsupervised model to cluster the input features into a lower dimensional output space. This output space would be the latent space W as mentioned above. Thus, similar feature vectors would have similar corresponding latent vectors. Autoencoders can also be used, where the latent space embeddings of the autoencoder can be directly used as the outputs of Phase 1.

3.2 Stage 2

The second phase of the Dual Phased Deep Learning Framework is a function which essentially converges pairs of embedded vectors, outputting the probability of a pair of vectors being semantically similar. This second phase is generally implemented using a Siamese Network along with a suitable distance metric like Euclidean distance. Both contrastive and triplet losses can be used to penalize the network into learning to differentiate between the feature embeddings. As a matter of fact, this framework itself is like a Siamese Network. To draw this analogy further, the initial layers of the Siamese Network are trained independently and frozen.

Since the h(x) and g(x) functions are trained independently and not at the same time, the requirement for GPU's memory is lower. The time required for training the overall model is also lesser.

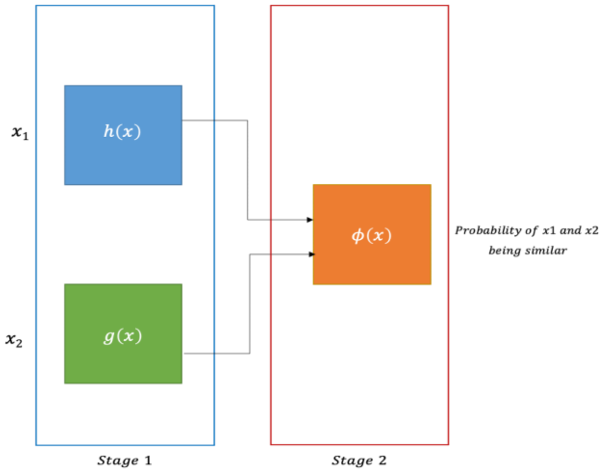


Fig -2: Proposed Dual Stage Framework

4. METHODOLOGY

For verifying the hypothesis proposed by the Dual Phased Framework, we perform experiments on Sketch Based Image Retrieval. The goal of this task is to retrieve the most relevant images from a set of images, given a hand drawn sketch. Relevancy here refers to the semantic similarity between the query sketch and the retrieved image.

Sketch Based Image Retrieval is a form of Content Based Retrieval, where the goal is to depict ideas in the form of sketches and find the most similar image from an image collection. Sketch Based Image Retrieval is inherently a Cross Domain problem [15], since we require a mapping between sketches and images, which are of different modalities. This domain gap between the sketch and its corresponding image can be bridged by using the proposed Dual Stage Framework.

In the first stage we extract the feature embedding of the sketches and images. This is done using Convolutional Neural Networks. This is followed by the second stage, where we train a model to converge these features i.e. the model will output a probability describing the similarity between the sketch and image embeddings. The following sections illustrate the process of training and testing in detail.

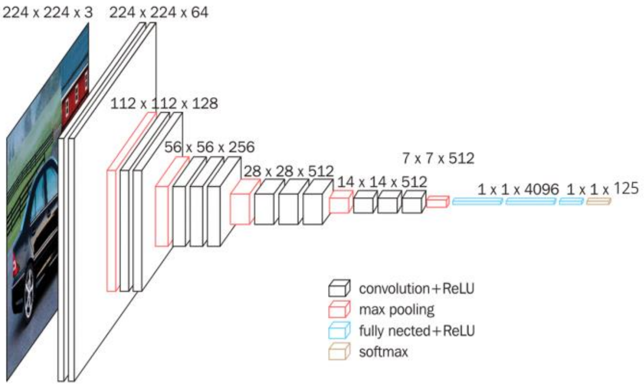


Fig -2: VGG16 Architecture

4.1 Dataset

In this work the sketchy database is used. It contains 75,471 sketches of 12,500 objects traversing 125 categories or classes. The Sketchy database defines relationships between pairs of images and sketches, and we use this mapping as a cross domain association.

The database consists of invalid and ambiguous sketches as well. These sketches are filtered out and the remaining sketches are split into train and test sets. Therefore, the set of sketches is split into 4 sets, including train, test, invalid and ambiguous sketches. The images corresponding to these sketches are also organized in a similar fashion. The mentioned statistics are depicted in Table 1.

Table -1: Dataset Metrics

Train	Test	Invalid	Ambiguous
58050	7332	885	9214

Train	Test
11209	1291

Although there is a significant imbalance with respect to the number of sketches and images, the proposed framework is robust to domain imbalance owing to the fact that pairs need not be formed, and the mapping functions are learned independently.

4.2 Model Architecture

The first phase consists of two VGG16 networks, which are Convolutional Neural Networks (CNN), which help us to extract the features from the sketches and images. The VGG16 network is modified such that the final layer is replaced by a fully connected layer having 125 nodes, one for each class. The features of the penultimate layer of the two CNN's, having 1024 nodes, are extracted to be used as the embedding vectors.

Once these features have been visualized, we pass them through the phase 2 network to train it. The second phase of the network consists of a Siamese Network. This network has 3 layers with the last layer being a fully connected layer with 125 neurons. The other layers consist of 1024 nodes each. Euclidean distance is used as a metric to determine similarity between the last layers of the Siamese Network.

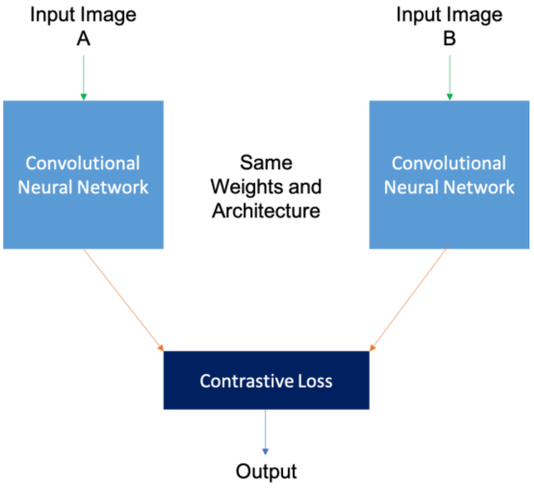


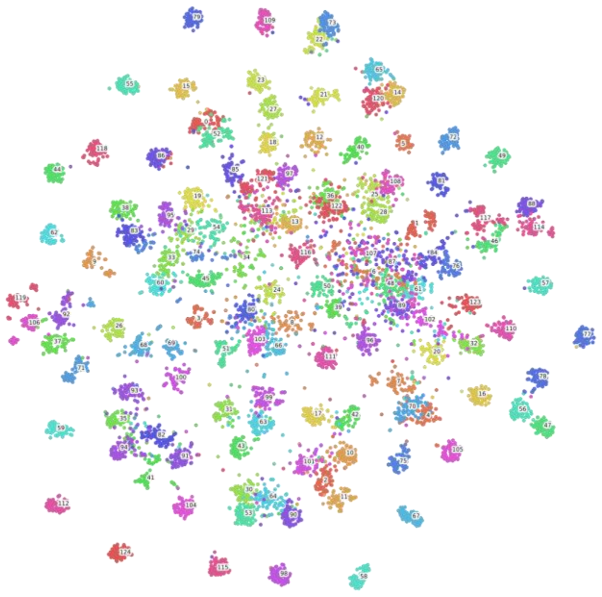
Fig -4: Siamese Network Architecture

4.1 Training

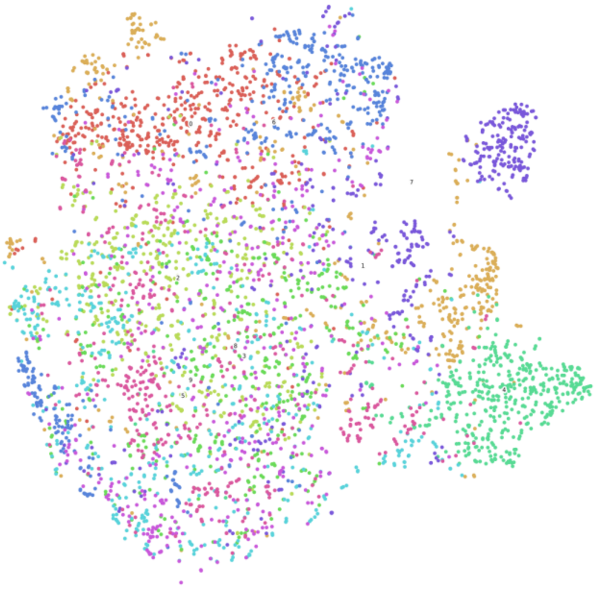
A transfer learning approach is used to train the two VGG16 models on sketches and images. These models are initially loaded with pre-trained weights and all layers except for the last four are frozen. These are then trained independently on the training set of images and sketches.

Both of these models are trained for 20 epochs after freezing all but the last layer. Then fine tuning is performed by unfreezing all the layers. The penultimate layer features are then extracted. These have a dimension of 1024, equal to the number of neurons in the penultimate layer.

These features intuitively represent the sketches and images projected onto 1024 dimensional space. Thus, the VGG16 networks act as dimensionality reduction mechanisms. Further, we visualize these using a dimensionality reduction technique called t-SNE [17] . t-SNE is used to visualize the features in higher dimensional latent space by projecting them onto 2 dimensional feature space using unsupervised learning techniques. These plots are shown in Figure 5. Next, we train the phase 2 of the model, the Siamese Network. The extracted features are paired in such a way that for a given sketch there are 10 pairs with images which are very similar or related, i.e., the positive pairs and there are 10 pairs with images which are dissimilar, i.e. the negative pairs. The Siamese Network is then trained using these pairs [18, 19].



Training images



Training sketches.

Fig -5: TSNE Embeddings

5. RESULTS AND OBSERVATIONS

For evaluating the retrieval results we use a metric called Mean Average Precision. Mean average precision is defined as the mean of average precision for a query. Average precision is calculated by plotting a precision recall curve and averaging the value of the continuous precision function. Mean average precision value [20] is considered for the retrieval of 100 images given a sketch query. This we calculate the mean of 100 average precisions.

We first present the results of the training on the vgg16 models. Table 2 shows the training and test accuracy, and Table 3 shows the comparison with baselines.

Table -2: Training metrics of Stage 1 models

Model	Training	Validation	Test
Vgg16 Sketch	99.84	71.14	69.17
Vgg16 Image	100	92.37	92.33

Table -3: Result Comparison

Model	MAP@100
Proposed Framework	**0.240**
Siamese Network end to end[1]	0.239
3D Shape[16]	0.192
GF-HOG[10]	0.122

As shown above, the model performs very well, and beats the existing benchmark scores for MAP@100 significantly. This proves that the model based on the Dual Stage Framework is very robust with respect to learning mappings between sketches and images.

6. CONCLUSIONS

In this paper, we proposed a novel Dual Stage Framework for Cross Domain Content Matching and Retrieval. There is significant reduction in the amount of resources required for training the cross domain network using this technique. Hence although the best performance may not be extracted, with limited demand for computational resources, using this framework can help produce strong results with limited GPU power within a short timeframe.

The hypothesis has been verified by training and testing a model on sketch-based image retrieval. A small Euclidean Distance is resultant from this model after being trained on positive and negative pairs of extracted latent vectors. This model outperforms the baselines and has validated the proposed framework. The Dual Stage Framework helps bridge the domain gap between dissimilar modalities. Future work can explore the use of this framework for any cross-domain problem given that this framework can generalize well to new modalities and domains while delivering robust performance.

REFERENCES

[1] Li Liu, , Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. "Deep Sketch Hashing: Fast Free-hand Sketch-Based Image Retrieval." (2017).
[2] Filip Radenovic and Giorgos Tolias and Ondrej Chum, . "Generic Sketch-Based Retrieval Learned without Drawing a Single Sketch".CoRR abs/1709.03409 (2017).
[3] Qingjie Meng, , Daniel Rueckert, and Bernhard Kainz. "Learning Cross-domain Generalizable Features by Representation Disentanglement." (2020).
[4] Shih, J.-L., Chen, L.-H., 2001. A new system for trademark segmentation and retrieval. in: Image Vision Computing, pp. 1011-1018.
[5] Sousa, P., Fonseca, M. J., 2010. Sketch-Based Retrieval of Drawings using Spatial Proximity , in: Journal of Visual Languages and Computing (JVLC) pp.69-80.
[6] Wang, C., Zhang, J., Yang, B., Zhang, L., 2011. Sketch2Cartoon: composing cartoon images by sketching, in: ACM Multimedia, pp. 789-790.
[7] Fonseca, M. J., Gonc¸alves, D., 2010. Sketch-a-Doc: Using Sketches to Find Documents, in: Proceedings of the 16th

International Conference on Distributed Multimedia Systems (DMS), pp. 327-330.

[8] Fonseca, M. J., Ferreira, A., Jorge, J. A., 2009. Sketch-based retrieval of complex drawings using hierarchical topology and geometry, in: Comput. Aided Des., pp. 1067-1081.

[9] Liang, S., Sun, Z., Li, B., 2005, Sketch Retrieval Based on Spatial Relations, in: CGIV, pp. 24-29.

[10] TRui Hu and John Collomosse. 2013. A performance evaluation of gradient field HOG descriptor for sketch based image retrieval. Comput. Vis. Image Underst. 117, 7 (July, 2013), 790–806. DOI:https://doi.org/10.1016/j.cviu.2013.02.005

[11] Y. Qi, Y. Song, H. Zhang and J. Liu, "Sketch-based image retrieval via Siamese convolutional neural network," 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 2460-2464, doi: 10.1109/ICIP.2016.7532801.

[12] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans. Graph. 35, 4, Article 119 (July 2016), 12 pages. DOI:https://doi.org/10.1145/2897824.2925954

[13] Meng, Qingjie, D. Rueckert and Bernhard Kainz. “Learning Cross-domain Generalizable Features by Representation Disentanglement.” ArXiv abs/2003.00321 (2020): n. Pag.

[14] H. Xia and Z. Ding, "Structure Preserving Generative Cross-Domain Learning," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4363-4372, doi: 10.1109/CVPR42600.2020.00442.

[15] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1386–1393, 2014

[16] W. Fang, K. Le, and L. Yi. Sketch-based 3d shape retrieval using convolutional neural networks. arXiv preprint arXiv:1504.03504, 2015.

[17] van der Maaten, Laurens and Hinton, Geoffrey. "Visualizing Data using t-SNE ." Journal of Machine Learning Research 9 (2008): 2579--2605

[18] Filip Radenović, Giorgos Tolias, & Ondřej Chum. (2018). Deep Shape Matching.

[19] Hongruixuan Chen and Chen Wu and Bo Du and Liangpei Zhang (2020). DSDANet: Deep Siamese Domain Adaptation Convolutional Neural Network for Cross-domain Change Detection. CoRR, abs/2006.09225.

[20] Kemal Oksuz, Baris Can Cam, Emre Akbas, & Sinan Kalkan. (2018). Localization Recall Precision (LRP): A New Performance Metric for Object Detection.