Domain Adaptation via Context Prediction for Engineering Diagram Search

Harsh Jhamtani^{1(✉)} and Taylor Berg-Kirkpatrick²

¹ Carnegie Mellon University, Pittsburgh, USA
jharsh@cs.cmu.edu

² UC San Diego, San Diego, USA
tberg@eng.ucsd.edu

Abstract. Effective search for engineering diagram images in larger collections is challenging because most existing feature extraction models are pre-trained on natural image data rather than diagrams. Surprisingly, we observe through experiments that even in-domain training with standard unsupervised representation learning techniques leads to poor results. We argue that, because of their structured nature, diagram images require more specially-tailored learning objectives. We propose a new method for unsupervised adaptation of out-of-domain feature extractors that asks the model to reason about spatial context. Specifically, we fine-tune a pre-trained image encoder by requiring it to correctly predict the relative orientation between pairs of nearby image regions. Experiments on the recently released Ikea Diagram Dataset show that our proposed method leads to substantial improvements on a downstream search task, more than doubling recall for certain query categories in the dataset.

Keywords: Diagram search · Image retrieval · Domain adaptation

1 Introduction

Many engineering enterprises maintain technical know-how about the design and working of various parts of their products via engineering diagrams. Such engineering diagrams typically show different parts used for a product or a portion of a product, and how these parts fit together in assemblies. In Fig. 1, the image on the left is a sample from an Ikea drawings dataset [1]. A common way of distributing such diagrams is in the form of images rather than detailed 3D models due to privacy issues and ease of use [1].

An important use case for such enterprises is the ability to automatically search for similar parts in diagram images based on a drawing or diagram image of a query part. A typical goal is to search a collection of engineering diagram images to find diagrams that contain the query part or a similar part as a component in the larger diagram, possibly at a different scales and rotations relative to the query. The matching diagram will likely contain many other parts that

are irrelevant to the query, which makes this task different from more standard image search scenarios where matched images will have the query as their central focus. In many ways, the most analogous task is that of matching a keyword query to a full-text document: matched keywords are usually surrounded by irrelevant text.

There have been only a few efforts directed towards search in engineering diagrams [1, 6]. Dai et al. [1] recently released a new Ikea dataset, which consists of 13,464 diagrams from Ikea manuals and 16,940 query images, and proposed a neural search engine for performing diagram search. The image features used by Dai et al. [1] are produced by a VGGNET model [16], pre-trained to predict object classes on ImageNet data [3]. Since ImageNet consists of natural images, rather than data like engineering diagrams, it is reasonable to expect that learning representations for diagrams using in-domain engineering data might lead to better results. Somewhat surprisingly, we show in experiments that tuning image representations using standard unsupervised techniques (e.g. using auto-encoding objectives) leads to worse performance than the out-of-domain pre-trained model. This is probably a result of structured nature of engineering diagrams that makes them visually distinct from natural images, on which the baseline unsupervised techniques were developed and validated [8].

In this paper we propose new unsupervised methods for tuning the pretrained model for engineering diagram data using spatial context as a learning signal. Specifically, a classifier trained to predict relative direction of two randomly sampled image patches provides the spatial signal. Use of spatial context alone to learn image representations *from scratch* has been explored in prior work [5], though not in the context of image search. However, we report that such an approach leads to poor results for engineering diagram search. We instead build on prior work to propose unsupervised *fine-tuning* of pre-trained neural image encoder models using the *spatial context*. Such use of spatial context to finetune the image encoder model has not been explored earlier.

In experiments, we find that the proposed approach leads to substantial gains in retrieval performance relative to past work as well as compared to a variety of unsupervised baselines, yielding state-of-the-art results on the Ikea dataset.

2 Related Work

Our work is related to Dai et al. [1] who introduce Ikea dataset and propose neural search methods for the task. However, they use pre-trained VGGNET model to extract image features and do not attempt learn or fine tune image feature extractors. Many earlier image retrieval methods are based on SIFT features and bag-of-word models [2, 9, 17]. Recently, convolutional and other neural models have been used more extensively as feature extractors for image retrieval [15, 20]. Use of region based image features has been found useful in some prior image retrieval work [14]. Our proposed method is also related to recent success on learning contextualized word representations [4, 10, 13] and image representations from scratch [5, 11] through self-supervised training by using context prediction as a training signal.

Fig. 1. Overview of the proposed method. For an image patch a , another patch b is sampled from one of the eight possible directions (shown with dotted borders). A classifier with parameters θ is trained to predict the relative direction d . The image encoder, with parameters ϕ , is biased to stay close to the original pre-trained parameters ϕ_0 via a L1-regularization term to encourage the model to retain useful features from the pre-trained model.

Most prior work on fine-tuning and domain adaptation for image representations requires in-domain supervised data, and either uses parameter fine-tuning [18, 19] or feature selection [7, 19] for domain adaptation. In contrast, our approach is a fully unsupervised method for domain adaptation of pre-trained models. Certain prior works on domain adaptation for image feature extractors require both the source and target domain data during training [19], which our approach does not require. Some prior works have explored using unsupervised methods like auto-encoders for domain adaptation [12].

3 Fine-Tuning by Context Prediction

We propose an unsupervised method that uses spatial context prediction as a learning signal to tune a pre-trained image feature extractor, which we refer to as an image encoder, for use in a neural engineering diagram search. Our approach extends prior work on context-based training objective [5] to perform domain adaptation and applies the method to a new domain and task. Specifically, our goal is to bias image encoders to capture more informative features of engineering diagrams by requiring them to predict spatial relationships between neighboring image regions. We shall refer to our fine-tuning method as **SPACES** (SPAtial Context for Engineering diagram Search).

Let $\mathcal{D}$ denote the set of images in the dataset. For an image $I \in \mathcal{D}$, we randomly pick a rectangular region in the image of size $M * M$ - an image patch (Fig. 1). Let us denote the identified patch by a . Thereafter, we choose one of the 8 cardinal directions (North, South, East, West, North-East, North-West, South-East, South-West), which we denote as d , uniformly at random. Then, a second rectangular patch b is identified close to the first patch in the sampled direction d such that b and a do not intersect. However, following [5], we identify

candidates for b a minimum fixed distance x from patch a , and introduce random jitters in horizontal and vertical directions. We observed $M = 24$ and $x = 4$ to be a reasonable choice. We denote the distribution from which a , b , and d are sampled as the generator, G .

We extract the features of the patches using the image encoder model, denoted by f_ϕ , where ϕ are the model parameters. f_ϕ is typically a deep convolutional neural network. Given a pre-trained image encoder model with parameter weights ϕ_0 , our task is to finetune the model using context prediction signal from a classifier defined as follows. A classifier with learnable parameters θ takes as input the extracted features of the two patches and makes a prediction about the relative direction of the patches. Specifically, we consider a two layer feed-forward neural network with a softmax function at the end to make a 8-way classification prediction (for 8 cardinal directions). Classification loss for a given patch pair a, b in relative direction d can be written as follows:

$$S(\phi, \theta) = \sum_{I \in \mathcal{D}} \mathbb{E}_{d, a, b \sim G(I)} [-\log(p_\theta(d | f_\phi(a), f_\phi(b)))] \quad (1)$$

Computing this exact loss is impractical due to an extremely large number of possible patch pairs. So we instead draw K random samples of pairs of patches for every image in the train set. Additionally, we regularize the image encoder model towards the pretrained model weights ϕ_0 by adding a L1 regularizer. This is done to encourage the model to retain many features from the pretrained model since abstract features like curves and shapes from have been shown to generalize well across tasks. We learn the image encoder and classifier jointly by optimizing for θ and ϕ to minimize the following loss function:

$$L(\phi, \theta) = \sum_{I \in \mathcal{D}} \frac{1}{K} \sum_{k=1}^K [-\log(p_\theta(d^{(k)} | f_\phi(a^{(k)}), f_\phi(b^{(k)})))] + \lambda |\phi - \phi_0| \quad (2)$$

The regularization term biases the image encoder parameters to remain closer to the original pre-trained model values. In early experiments, we observe that using L1 for this term performs better than using a L2 version. The classification loss term is based on the work of Doersch et al. [5]. However, we use the spatial context loss to fine-tune a pre-trained image feature extractor for the target engineering diagrams domain. In contrast, Doersch et al. [5] learn image representations from scratch using a large dataset of natural images and focus on a different task. We demonstrate in experiments that the proposed tuning method substantially outperforms the training from scratch for our domain and task.

4 Experiments and Results

4.1 Dataset

We use the Ikea dataset [1] which consists of 13,464 furniture assembly diagrams. Each assembly diagram is a black-and-white image, and resembles a line

drawing. Query images are generated automatically from a subset of documents using an iterative procedure proposed in past work [1]. The procedure begins with identifying a localized region of high density black pixels, and keeps on expanding it until the black pixel density is lower than a threshold. The Ikea dataset consists of 5 query types: *psr*, *Psr*, *pSr*, *psR*, *PSR*: Lowercase letters *p*, *s*, *r* signify that position, scale and rotation, respectively, are unchanged in the generated query relative to the original image from which the query was extracted. Capital letters denote the corresponding altered attribute. Thus, for ***psr*** queries set, the identified region is placed onto a white background image of size same as original image, and at same position as the identified region was in the original image. ***psR*** queries are constructed by rotating *psr* queries, ***pSr*** queries are constructed by scale transformations, and so on.

Table 1. Retrieval results on different query types in Ikea dataset.

Model	Invariant (<i>psr</i>)		Position (<i>Psr</i>)		Scale (<i>pSr</i>)		Rotation (<i>psR</i>)		ALL (<i>PSR</i>)	
	MRR	R@1	MRR	R@1	MRR	R@1	M-RR	R@1	MRR	R@1
VGG [1]	0.94	0.89	0.90	0.84	0.78	0.70	0.38	0.28	0.17	0.10
VGG-AE	0.93	0.89	0.89	0.84	0.76	0.69	0.41	0.30	0.16	0.08
CTXT [5]	0.83	0.78	0.54	0.45	0.02	0.01	0.15	0.08	0.03	0.0
SPACES	0.98	0.98	0.96	0.95	0.88	0.84	0.65	0.58	0.22	0.14
SPACES-L	0.95	0.90	0.91	0.86	0.87	0.81	0.61	0.52	0.21	0.14

4.2 Experiment Setup

We use our proposed method to fine-tune VGGNET [16] (a deep convolutional image encoder, pre-trained on ImageNet [3] data). We report recall and mean reciprocal rank in downstream search, using the DISHCONV [1], a neural retrieval method which utilizes pairwise training over features extracted from convolutional kernels over image representations. We perform early stopping during training based on *recall@1* for queries in the validation split. We consider following baselines: (1) **VGG** represents a fixed pre-trained VGGNET model, trained on ImageNet data, as used in prior work [1]. (2) **VGG-AE** fine-tunes a pre-trained VGGNET using an autoencoder (with a deconvolutional network as decoder) with reconstruction objective. (3) **CTXT** involves training image representations from scratch just using context prediction [5]. We consider VGGNET architecture (initialized randomly) to encode images, and trained to predict relative direction of pairs of image patches.

4.3 Results

Table 1 summarizes the results when evaluating downstream search performance. Overall, SPACES performs much better than the baselines VGG, VGG-AE and CTXT across different query types. Recall that *psr* are the most basic

Fig. 2. MRR (Mean Reciprocal rank) plotted against (a) the scale transformation value of queries in pSr test set (b) rotation degrees in psR test set.

query types, *psR* are queries created by rotating basic *psr* queries by varying degrees, and so on. The largest improvement over baselines is observed for *pSr* and *psR* query types. SPACES performs better than VGG-AE model probably because it has a more suitable training signal given the structured nature of the engineering diagrams. The baseline CTXT model has to learn the image encoder model entirely from a relatively small number of images. In contrast, SPACES leverages the pre-trained model and is able to fine-tune on the Ikea dataset using only a few thousand images. This demonstrates the utility of SPACES in adapting large pre-trained image encoder models for engineering diagrams.

We report MRR for a range scale factor and rotations degrees (Fig. 2). SPACES performs better than the baseline almost all throughout different scale and rotation changes. We also report the results with SPACES-L which uses total loss on the validation split for early stopping instead of recall@1 scores. MRR and recall scores from these variants are observed to be very similar (Table 1) demonstrating that the proposed approach is robust to such changes in early stopping criteria.

A relative direction prediction classifier trained on the Ikea dataset images with features from pretrained VGGNET model, and then evaluated on 1000 patch pair samples achieves only 14.2% accuracy, which is close to performance of a random prediction classifier for a 8-way classification problem. The trained classifier within SPACES achieved 43% accuracy in the 8-way classification task, which demonstrates that features from our trained model encode more information about neighboring context.

5 Conclusions

In this paper we have proposed an unsupervised method to adapt a pre-trained neural image encoder on an engineering diagram dataset using spatial context prediction. We demonstrate that standard unsupervised representation learning methods such as autoencoder are not amenable to engineering diagrams, probably due to their structured nature. Our proposed method outperforms the original pre-trained feature extractor as well as other unsupervised baselines to achieve state-of-the-art results on Ikea dataset.

References

1. Dai, Z., Fan, Z., Rahman, H., Callan, J.: Local matching networks for engineering diagram search. In: The World Wide Web Conference, WWW 2019 (2019)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: ideas, influences, and trends of the new age. *ACM Comput. Surv. (CSUR)* **40**(2), 5 (2008)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)* (2018)
5. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1422–1430 (2015)
6. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: benchmark and bag-of-features descriptors. *IEEE Trans. Vis. Comput. Graph.* **17**(11), 1624–1636 (2010)
7. Krause, J., Jin, H., Yang, J., Fei-Fei, L.: Fine-grained recognition without part annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5546–5555 (2015)
8. Krizhevsky, A., Hinton, G.E.: Using very deep autoencoders for content-based image retrieval. In: ESANN (2011)
9. Lowe, D.G., et al.: Object recognition from local scale-invariant features. In: ICCV 1999, pp. 1150–1157 (1999)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
11. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving Jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
12. Parchami, M., Bashbaghi, S., Granger, E., Sayed, S.: Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
13. Peters, M.E., et al.: Deep contextualized word representations. In: Proceedings of NAACL-HLT, pp. 2227–2237 (2018)
14. Pham, T.T., Maillot, N.E., Lim, J.H., Chevallet, J.P.: Latent semantic fusion model for image retrieval and annotation. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 439–444. ACM (2007)
15. Razavian, A.S., Sullivan, J., Carlsson, S., Maki, A.: Visual instance retrieval with deep convolutional networks. *ITE Trans. Media Technol. Appl.* **4**(3), 251–258 (2016)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)* (2014)

17. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, ICCV 2003, vol. 2, p. 1470. IEEE Computer Society, USA (2003)
18. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. Imaging* **35**(5), 1299–1312 (2016)
19. Wang, M., Deng, W.: Deep visual domain adaptation: a survey. *Neurocomputing* **312**, 135–153 (2018)
20. Zhou, W., Li, H., Tian, Q.: Recent advance in content-based image retrieval: a literature survey. arXiv preprint [arXiv:1706.06064](https://arxiv.org/abs/1706.06064) (2017)