Unsupervised Sketch-Based Image Retreival using Cross-Domain Context Prediction

Final Presentation

Anant Joshi, Osama Sakhi, Sarmishta Velury

November 26, 2019

Georgia Institute of Technology

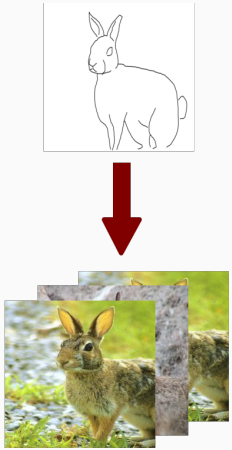
Table of Contents

Problem Statement

Problem Statement

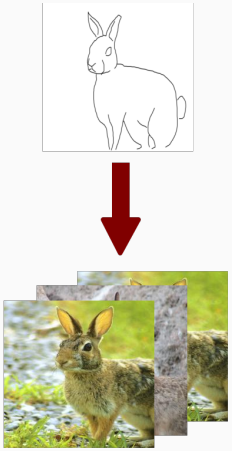
- Given a hand-drawn sketch.

Problem Statement



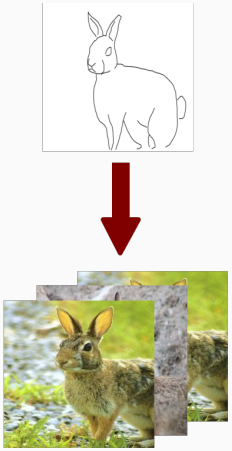
- Given a hand-drawn sketch.
- Retrieve the best matching images from a dataset.

Problem Statement: Detail



- We intend to solve **unsupervised** fine-grained Sketch-Based Image Retrieval.

Problem Statement: Detail

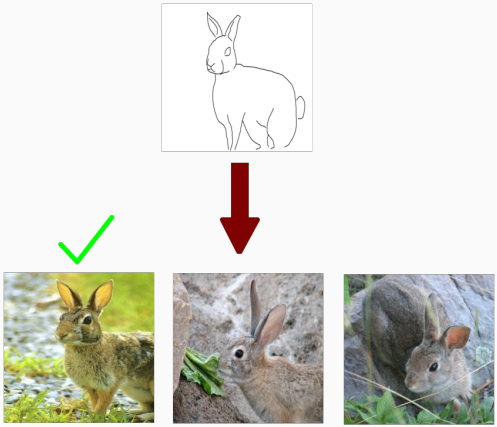


- We intend to solve **unsupervised** fine-grained Sketch-Based Image Retrieval.
- This means, we retrieve **specific** instances of entities, such as a plump bunny with pointy ears, resting on its forelegs, and facing left.

Problem Statement: More Detail

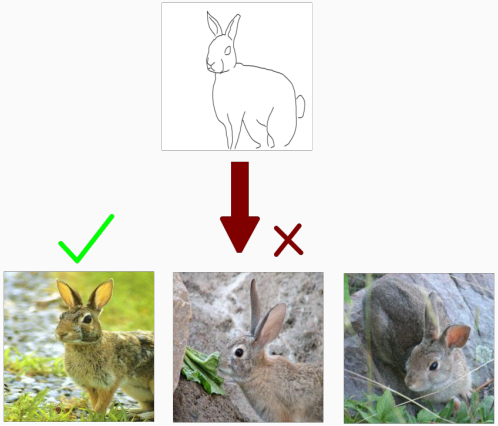
- Here, we can see a set of 3 retrievals.

Problem Statement: More Detail



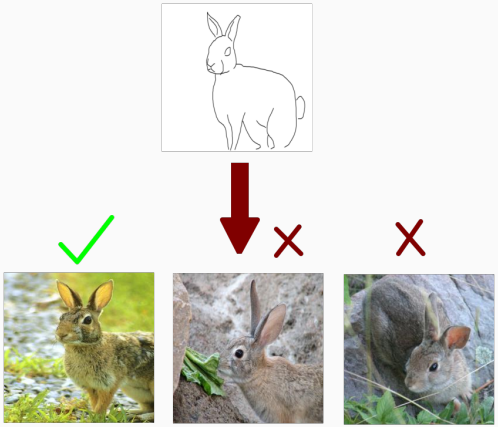
- Here, we can see a set of 3 retrievals.
- The first is correct, as the appearance and pose are correct.

Problem Statement: More Detail



- Here, we can see a set of 3 retrievals.
- The first is correct, as the appearance and pose are correct.
- In the second, the orientation is similar, but not quite correct.

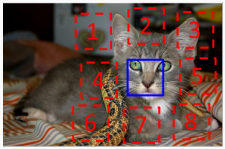
Problem Statement: More Detail



- Here, we can see a set of 3 retrievals.
- The first is correct, as the appearance and pose are correct.
- In the second, the orientation is similar, but not quite correct.
- In the third, the orientation and pose are completely wrong.

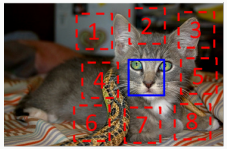
Table of Contents

Related Work: Unsupervised Visual Representation Learning by Context Prediction(1)



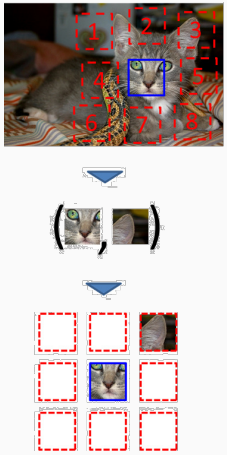
- **Task:** Context prediction by spatially relating two random patches of an image.

Related Work: Unsupervised Visual Representation Learning by Context Prediction(1)



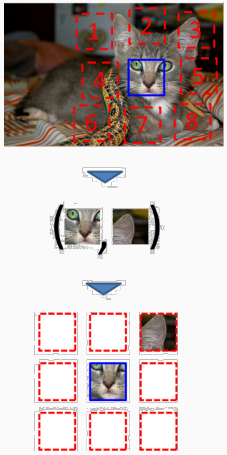
- **Task:** Context prediction by spatially relating two random patches of an image.

Related Work: Unsupervised Visual Representation Learning by Context Prediction(1)



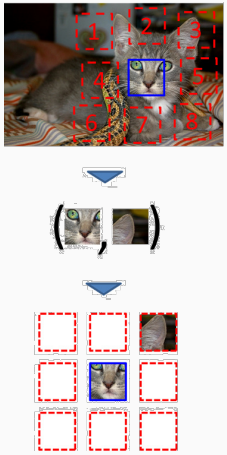
- **Task:** Context prediction by spatially relating two random patches of an image.

Related Work: Unsupervised Visual Representation Learning by Context Prediction(1)



- **Task:** Context prediction by spatially relating two random patches of an image.
- **Motivation:** Learn a feature embedding for images, such that images which are visually similar would be close in the embedding space.

Related Work: Unsupervised Visual Representation Learning by Context Prediction(1)

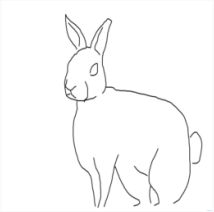


- **Task:** Context prediction by spatially relating two random patches of an image.
- **Motivation:** Learn a feature embedding for images, such that images which are visually similar would be close in the embedding space.
- **Method:** Late fusion architecture using two AlexNet style architectures, fused at fc6.

Related Work: The Sketchy Database

- **Task:** Create a large collection of paired image-sketch data for fine-grained sketch-based image retrieval.

Related Work: The Sketchy Database



- **Task:** Create a large collection of paired image-sketch data for fine-grained sketch-based image retrieval.
- **Motivation:** Large and detailed corpus of individual entities across domains, along with baseline benchmarks.

Related Work: Cross-modal Subspace Learning for fine-grained sketch-based image retrieval(4)

- **Task:** Introduce and compare a series of SOTA cross-domain subspace learning methods.

Related Work: Cross-modal Subspace Learning for fine-grained sketch-based image retrieval(4)

- **Task:** Introduce and compare a series of SOTA cross-domain subspace learning methods.
- **Motivation:** Study the effectiveness of cross-modal matching methods for image and text in SBIR.

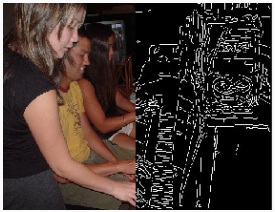
Related Work: Cross-modal Subspace Learning for fine-grained sketch-based image retrieval(4)

- **Task:** Introduce and compare a series of SOTA cross-domain subspace learning methods.
- **Motivation:** Study the effectiveness of cross-modal matching methods for image and text in SBIR.
- **Conclusion:** Demonstrate empirically that subspace learning can bridge the image-sketch domain gap.

Novelty

- **Unsupervised:** Most sketch-based image retrieval tasks are often fully supervised, and specialized for the task. Our approach aims to be an unsupervised approach to reconciling the domain gap using cross-domain context prediction.

Novelty



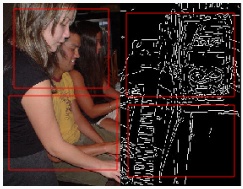
- **Unsupervised:** Most sketch-based image retrieval tasks are often fully supervised, and specialized for the task. Our approach aims to be an unsupervised approach to reconciling the domain gap using cross-domain context prediction.
- **Cross-domain:** Unlike the original paper by Doersch(1), our context encoder is trained across domains to make it learn domain-invariant features.

Table of Contents

Approach: Assumptions

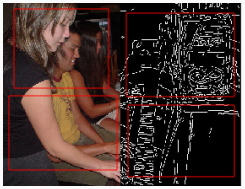
- Aligned, paired images available. For this, we compute canny edges of the images in the PASCAL VOC dataset.
- Clustering image and sketch embeddings from a well-trained network will result in well-formed discrete clusters that are domain agnostic.
- The model that performs well on cross domain context prediction will perform well on the cross-domain image retrieval task.

Pretext Task



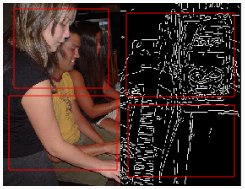
- We divide the image into 4 regions, with uneven spacing and jitter.

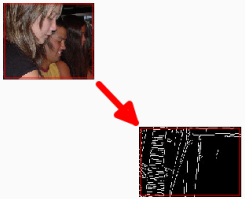
Pretext Task



- We divide the image into 4 regions, with uneven spacing and jitter.
- We then extract two patches, one from each domain, i.e. Images from Pascal, and their Canny edges.

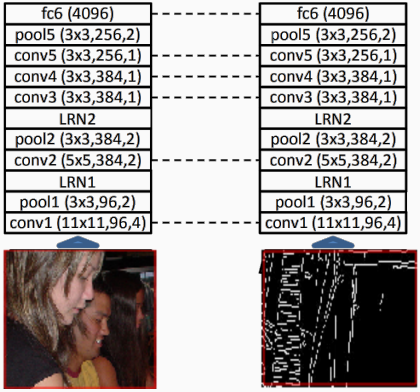
Pretext Task





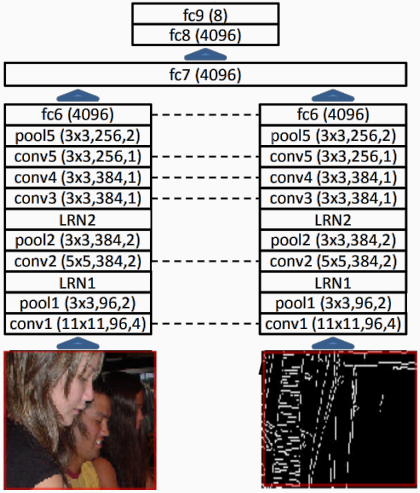
- We divide the image into 4 regions, with uneven spacing and jitter.
- We then extract two patches, one from each domain, i.e. Images from Pascal, and their Canny edges.
- We finally compute the relative positioning of the patches using the context encoder.

Pretext Task: Architecture



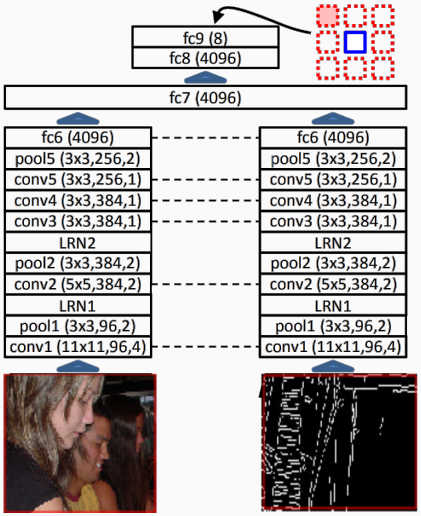
- We pass the two patches through the AlexNet model, with joint features.

Pretext Task: Architecture



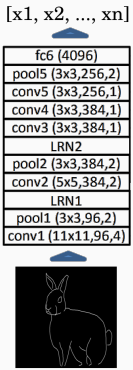
- We pass the two patches through the AlexNet model, with joint features.
- The two outputs are then concatenated into one joint embedding, and passed through more fully-connected layers.

Pretext Task: Architecture



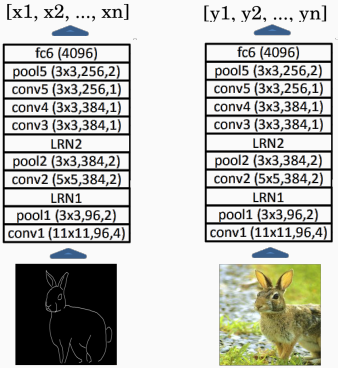
- We pass the two patches through the AlexNet model, with joint features.
- The two outputs are then concatenated into one joint embedding, and passed through more fully-connected layers.
- The relative 8-way position is classified, and the errors are backpropagated.

Image Retrieval



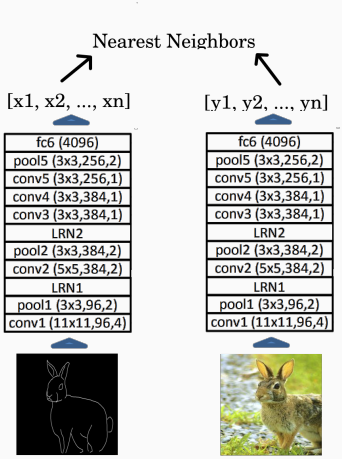
- We first compute embeddings for the query sketch using AlexNet trained on the pretext.

Image Retrieval



- We first compute embeddings for the query sketch using AlexNet trained on the pretext.
- We then perform a nearest neighbour search on the embeddings from the dataset of images.

Image Retrieval



- We first compute embeddings for the query sketch using AlexNet trained on the pretext.
- We then perform a nearest neighbour search on the embeddings from the dataset of images.
- We retrieve the nearest 5 and 10 images for top-5 and top-10 scores.

Contribution

- We extend the context-encoder concept across multiple domains.

Contribution

- We extend the context-encoder concept across multiple domains.
- We use the obtained embeddings to perform **unsupervised** fine-grained Sketch Based Image Retrieval.

Table of Contents

Datasets

1. **PASCAL VOC 2012 (2)**: We use this dataset for training our model on the pretext task.

Datasets

1. **PASCAL VOC 2012 (2)**: We use this dataset for training our model on the pretext task.
2. **Canny Edges of PASCAL VOC 2012**: We generate Canny Edges of the PASCAL VOC dataset using OpenCV's implementation of the Canny Edge Detection algorithm. We use this generated dataset for training on our pretext task.

Datasets

1. **PASCAL VOC 2012 (2)**: We use this dataset for training our model on the pretext task.
2. **Canny Edges of PASCAL VOC 2012**: We generate Canny Edges of the PASCAL VOC dataset using OpenCV's implementation of the Canny Edge Detection algorithm. We use this generated dataset for training on our pretext task.
3. **Sketchy (3)**: A large-scale collection of image-sketch pairs collected using Amazon's Mechanical Turk. We use this dataset for fine-grained SBIR.

Experiments

For all our experiments, we use the AlexNet architecture. We train each model for 26 epochs, with decreasing levels of ImageNet pretraining.

Experiments

For all our experiments, we use the AlexNet architecture. We train each model for 26 epochs, with decreasing levels of ImageNet pretraining.

- **Context Prediction in one domain**: We perform this task on PASCAL images and the generated Canny Edges individually.

Experiments

For all our experiments, we use the AlexNet architecture. We train each model for 26 epochs, with decreasing levels of ImageNet pretraining.

- **Context Prediction in one domain**: We perform this task on PASCAL images and the generated Canny Edges individually.
- **Context Prediction in two domains**: We extract a patch each from a PASCAL images and its corresponding edge image. The model is then trained to position them spatially.

Experiments

For all our experiments, we use the AlexNet architecture. We train each model for 26 epochs, with decreasing levels of ImageNet pretraining.

- **Context Prediction in one domain**: We perform this task on PASCAL images and the generated Canny Edges individually.
- **Context Prediction in two domains**: We extract a patch each from a PASCAL images and its corresponding edge image. The model is then trained to position them spatially.
- **Sketch Based Image Retrieval**: We use our model trained on the pretext task to extract features from the query sketch and perform a nearest neighbours search on our dataset to find the matching image.

Baselines

- **Pretrained AlexNet:** We compare performance on the Image Retrieval task using AlexNet pre-trained on ImageNet with no additional training or fine-tuning.

Baselines

- **Pretrained AlexNet:** We compare performance on the Image Retrieval task using AlexNet pre-trained on ImageNet with no additional training or fine-tuning.
- **Siamese Network + Triplet Loss:** We compare our performance against the supervised method defined in the Sketchy(3) paper.

Baselines

- **Pretrained AlexNet:** We compare performance on the Image Retrieval task using AlexNet pre-trained on ImageNet with no additional training or fine-tuning.
- **Siamese Network + Triplet Loss:** We compare our performance against the supervised method defined in the Sketchy(3) paper.
- **Li et. al.(5):** We compare our scores with Li et. al.'s supervised method (Deformable Part-based Model (DPM)), as done in the Sketchy(3).

Baselines

- **Pretrained AlexNet:** We compare performance on the Image Retrieval task using AlexNet pre-trained on ImageNet with no additional training or fine-tuning.
- **Siamese Network + Triplet Loss:** We compare our performance against the supervised method defined in the Sketchy(3) paper.
- **Li et. al.(5):** We compare our scores with Li et. al.'s supervised method (Deformable Part-based Model (DPM)), as done in the Sketchy(3).
- **Spatial Pyramid:** This model provides an improvement over traditional BOW models and has been used as a baseline in many works.

Metrics

We use two metrics to compare performance on the Image Retrieval task:

- top-5 accuracy
- top-10 accuracy

We quantify performance on the context prediction task using accuracy of relative position prediction.

Code & Implementation

- We used PyTorch to recreate Doersch's (1) Context Prediction Network.
- We use nearest neighbours to retrieve images for a given sketch in Sketchy(3).

Quantitative Results

Top 5	AlexNet Pretrained	Ours*	Li	Sketchy	SP
airplane	15.94	16.36	22.0	**27.2**	20.33
bicycle	6.68	8.79	11.67	**21.5**	13.83
car	10.90	11.99	**18.83**	15.8	14.5
cat	12.28	13.73	12.17	**13.8**	7.67
chair	17.49	13.30	20	**21.7**	20.33
cow	13.15	12.73	19.67	**19.8**	14
dog	16.04	13.15	9.5	**21**	6.83
horse	11.51	12.87	**31.67**	23.2	7.33
motorcycle	9.64	9.95	**22.5**	13	9
sheep	11.66	15.84	17.67	**21**	5
Mean	12.53	12.87	18.57	**19.8**	11.88

*We report results and analyses using the full pretrained + Batch Norm model.

Table of Contents

Ablations: Variables

We experiment with the following variables for our ablations.

Ablations: Variables

We experiment with the following variables for our ablations.

- **Pretraining:** To accelerate our training, we initialize the model with pretrained Imagenet weights. We vary the levels of pretraining to see how it affects performance.

Ablations: Variables

We experiment with the following variables for our ablations.

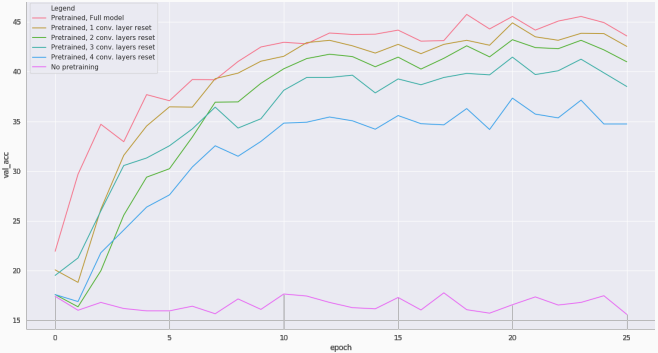
- **Pretraining:** To accelerate our training, we initialize the model with pretrained Imagenet weights. We vary the levels of pretraining to see how it affects performance.
- **BatchNorm:** Doersch et. al.(1) use BatchNorm to improve their performance.

Ablations: Variables

We experiment with the following variables for our ablations.

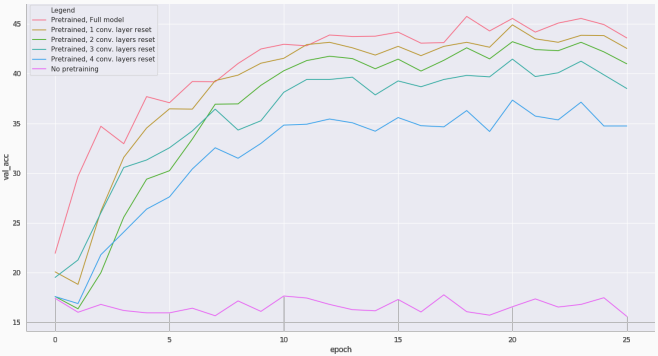
- **Pretraining:** To accelerate our training, we initialize the model with pretrained Imagenet weights. We vary the levels of pretraining to see how it affects performance.
- **BatchNorm:** Doersch et. al.(1) use BatchNorm to improve their performance.
- **Domains:** We see how the context encoder method performs when trained on each domain individually, as well as across domains.

Ablations: Pretraining (Validation accuracy on Cross Domain



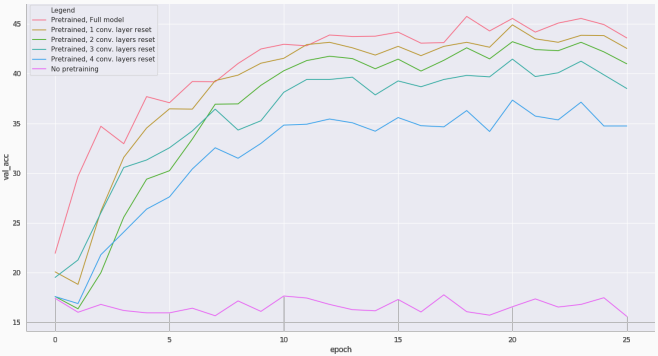
- NoPretrain has low accuracy, and doesn't learn much.

Ablations: Pretraining (Validation accuracy on Cross Domain



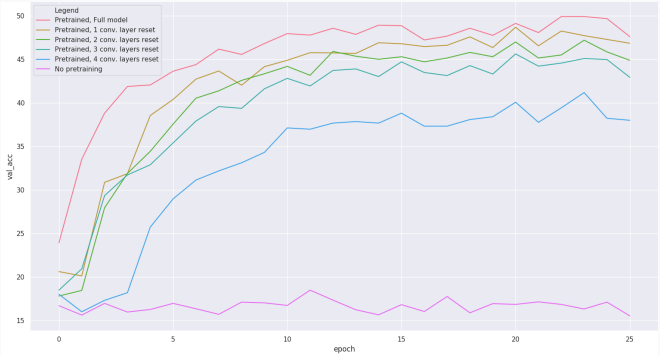
- NoPretrain has low accuracy, and doesn't learn much.
- Full pretrain has a very high validation accuracy.

Ablations: Pretraining (Validation accuracy on Cross Domain



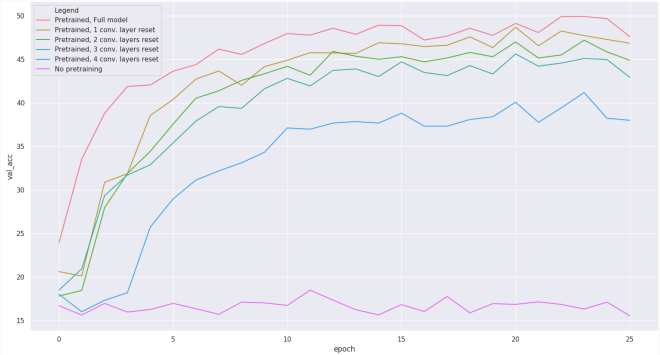
- NoPretrain has low accuracy, and doesn't learn much.
- Full pretrain has a very high validation accuracy.
- Resetting higher layers corresponds to small 2-3% drops in accuracy, which shows that our model needed lower-level features to jumpstart training.

Ablations: Single Domain — Pascal Images



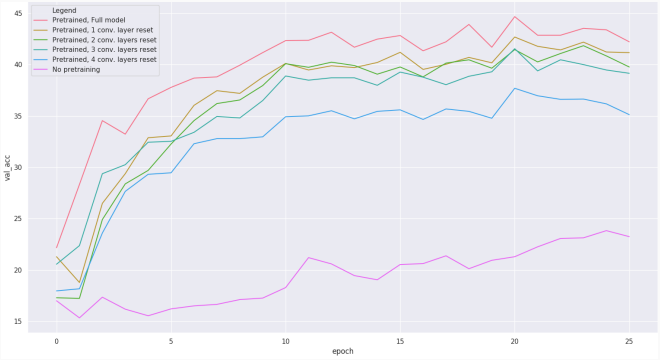
- This shows similar characteristics to cross-domain training.

Ablations: Single Domain — Pascal Images



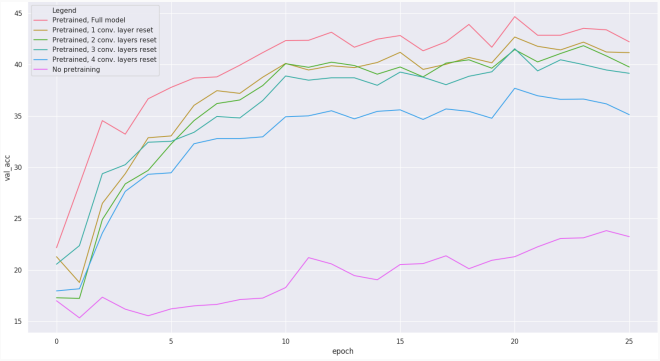
- This shows similar characteristics to cross-domain training.
- We can see how even the lowest pretrained convolution layer massively helps our training.

Ablations: Single Domain — Canny Images



- This also shows similar characteristics to cross-domain training.

Ablations: Single Domain — Canny Images



- This also shows similar characteristics to cross-domain training.
- However, since the low-level features are relatively simple, even NoPretrain begins to learn somewhat.

Ablations: Cross Domain — BatchNorm

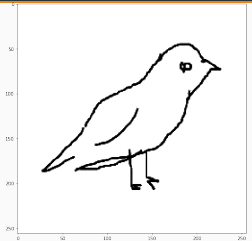
- Here, we can see that using BatchNorm, NoPretrain is able to escape the saddle point.

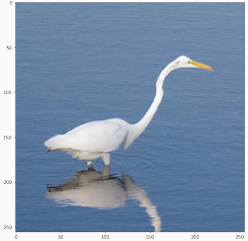
Ablations: Cross Domain — BatchNorm

- Here, we can see that using BatchNorm, NoPretrain is able to escape the saddle point.
- However, the model performance is significantly reduced with higher levels of pretraining.

Visualisations: Qualitative Samples — Good Result

(Results are drawn randomly)

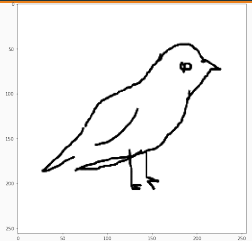


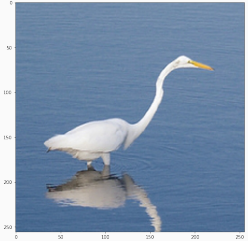


- Here, we can see that the intended image is present in the retrieved samples.

Visualisations: Qualitative Samples — Good Result

(Results are drawn randomly)

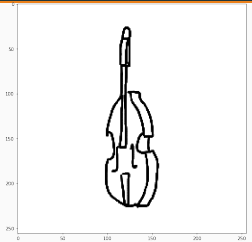


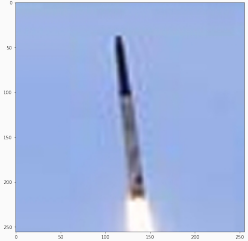


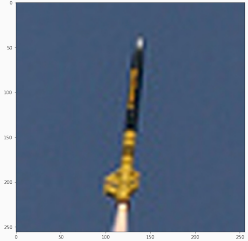
- Here, we can see that the intended image is present in the retrieved samples.
- We can also see that another result is also a bird in a similar pose.

Visualisations: Qualitative Samples — Bad Result

(Results are drawn randomly)



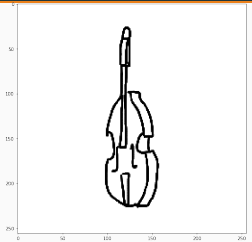


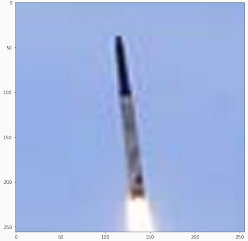


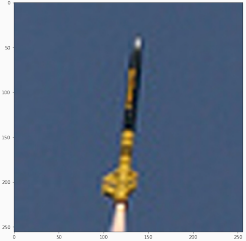
- Here, the intended class, i.e. the violin is not present.

Visualisations: Qualitative Samples — Bad Result

(Results are drawn randomly)

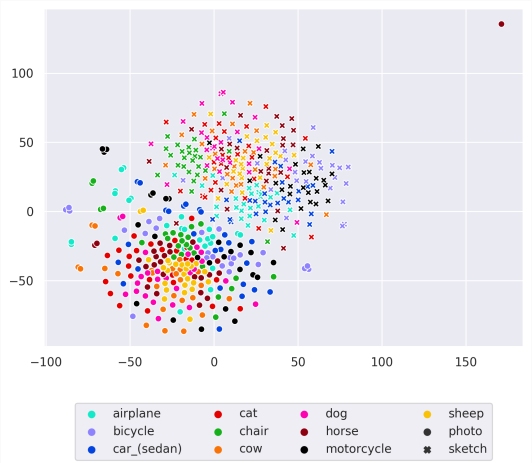






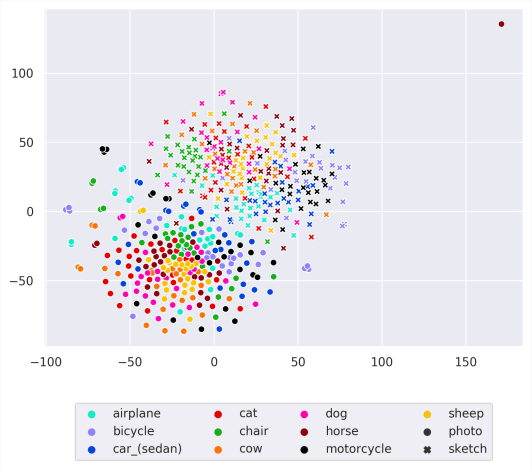
- Here, the intended class, i.e. the violin is not present.
- The query results make some sense as they are also long and thin objects, kept vertically.

Analysis: t-SNE — Sketchy (AlexNet Pretrained)



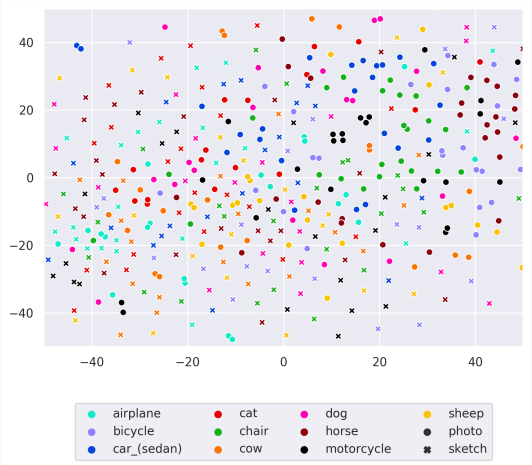
- Here, we can see there is a clear separation between the embeddings from the two domains.

Analysis: t-SNE — Sketchy (AlexNet Pretrained)



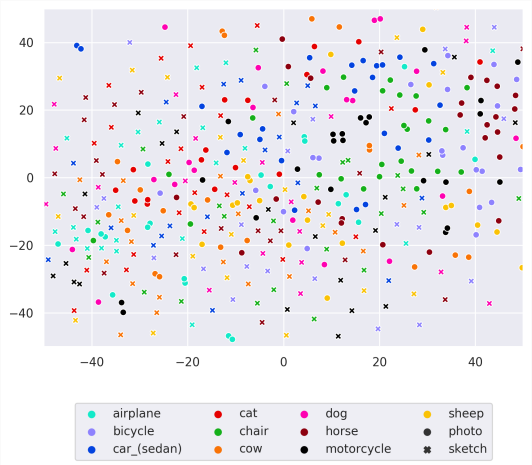
- Here, we can see there is a clear separation between the embeddings from the two domains.
- Corresponding classes are also pretty far from each other.

Analysis: t-SNE — Sketchy (Ours)



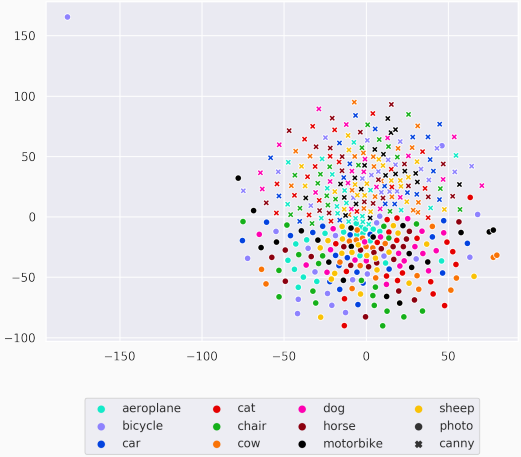
- Here, we can see the domains are much less separated. We can also see that in some classes, the distance between corresponding instances is getting lower.

Analysis: t-SNE — Sketchy (Ours)



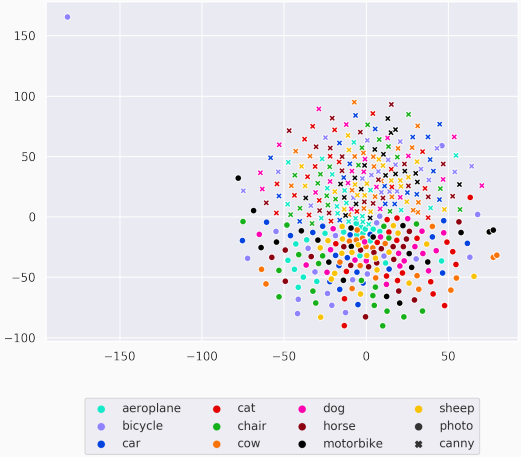
- Here, we can see the domains are much less separated. We can also see that in some classes, the distance between corresponding instances is getting lower.
- Overall, the results look better.

Analysis: t-SNE — Pascal ↔ Canny (AlexNet pretrained)



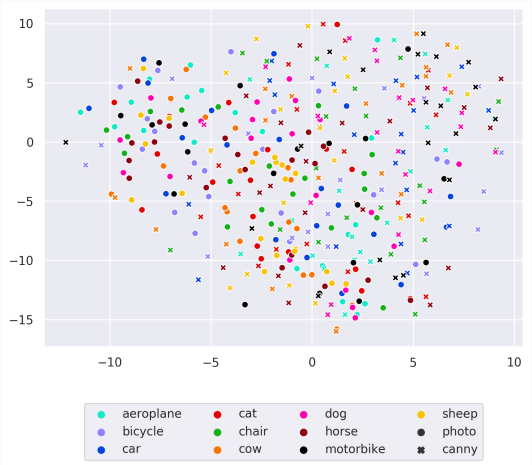
- Even here, there is a clear separation between the domains.

Analysis: t-SNE — Pascal ↔ Canny (AlexNet pretrained)



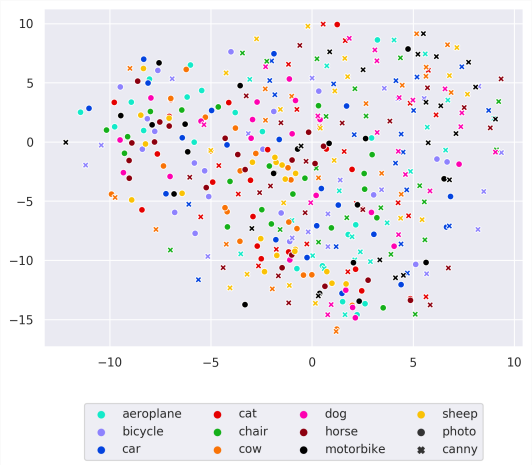
- Even here, there is a clear separation between the domains.
- The distance between corresponding points is high as well.

Analysis: t-SNE — Pascal ↔ Canny (Ours)



- This shows even better results, with good mixing in the domains, and close correspondences between images and sketches for some samples

Analysis: t-SNE — Pascal ↔ Canny (Ours)



- This shows even better results, with good mixing in the domains, and close correspondences between images and sketches for some samples
- It also shows that some classes are beginning to come closer. We think it means that we may need more training.

Table of Contents

Conclusion

- Our model performs better than the supervised classical baseline (Spatial Pyramid), and slightly better than a pretrained AlexNet.
- It shows better domain invariance compared to the AlexNet baseline.

Future Work

- Effects of further training on pretext task.

Future Work

- Effects of further training on pretext task.
- Use context-encoder as pretraining for supervised image retreival models.

Future Work

- Effects of further training on pretext task.
- Use context-encoder as pretraining for supervised image retreival models.
- Use more sophisticated feature extractors (like GoogLeNet or VGG) that more recent SBIR methods use.

References i

[1] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[2] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[3] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 2016.

References ii

[4] P. Xu, Q. Yin, Y. Huang, Y.-Z. Song, Z. Ma, L. Wang, T. Xiang, W. B. Kleijn, and J. Guo. Cross-modal subspace learning for fine-grained sketch-based image retrieval. *Neurocomputing*, 278:75–86, 2018.

[5] Y.-Z. S. S. G. Yi Li, Timothy M. Hospedales. Fine-grained sketch-based image retrieval by matching deformable part models. 2014.