General Domain Neural Networks for Specific Image Similarity Domain

Bachelor Oral Thesis Defence

Peter Friborg Jacobsen – wtk866
Supervisor: Desmond Elliot

UNIVERSITY OF COPENHAGEN

8th February - 2021

Introduction

Topic

- Transfer learning used to teach in-domain features to a general-domain CNN

Question

- Is it possible to close the gap between a general-domain CNN and an in-domain CNN, using only limited in-domain data?

Objective

- Provide a possible solution for small companies, through machine learning and opensource systems, to capitalise on their limited datasets

In-domain and general-domain

In-domain CNN

- Previous study[1] showed that an in-domain CNN would outperform the general-domain CNN by 10% points. The in-domain CNN were pretrained on the in-domain Places365 dataset.

General-domain CNN

- The general-domain CNNs proposed in this, were built on the ResNet50 CNN and EffecientNetB0, both pretrained on the general-domain ImageNet dataset.

In-domain dataset

- A limited in-domain dataset were studied in this thesis, along with the previous study. The in-domain dataset is a small selection of images from the Danish housing market, consisting of 6500 labeled images, and 14500 unlabeled images.

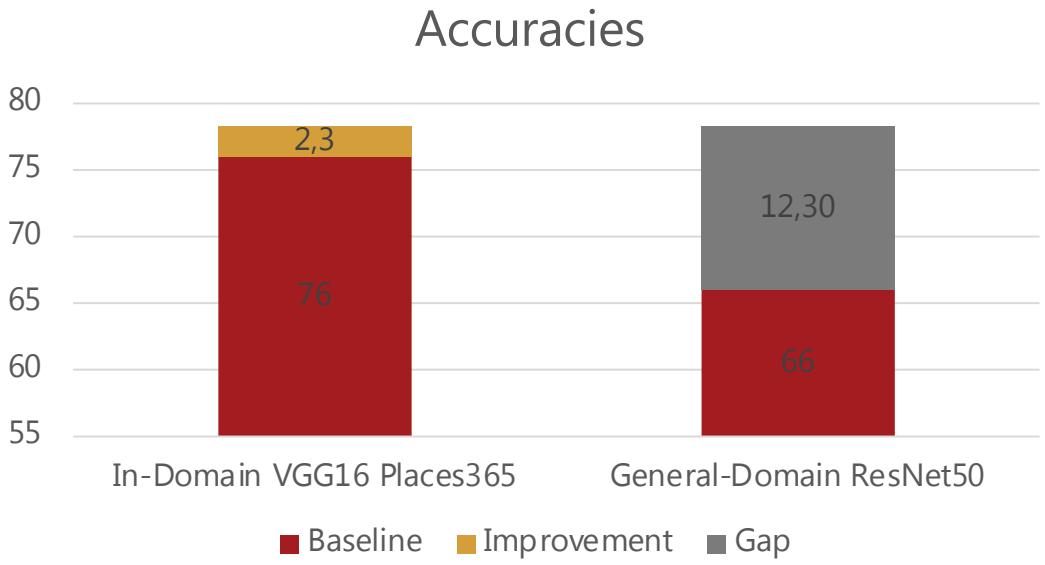
1. https://github.com/adamingwersen/ dre19_report/blob/master/master.pdf

Previous Work

Ingwersen: In-Domain

Ingwersen proposed a CNN, pretrained on the in-domain Places365 dataset.

The baseline of the in-domain were obtained to be superior to the general-domain model.



- An inconsistency were discovered when recreating the results from Ingwersens study regarding the accuracy of the baseline of the general-domain model

Khosla: Supervised Contrastive Learning

Khosla introduces a variant of contrastive learning, which takes advantage of the presence of labels.

The contrastive model introduced in Khosla is outperforming the previous state of the art classifier for various different datasets.

- Including the classification accuracy for the ImageNet dataset. Which the general-domain model is pretrained on

The top performing ResNet contrastive model introduced in the study improved the accuracies by:

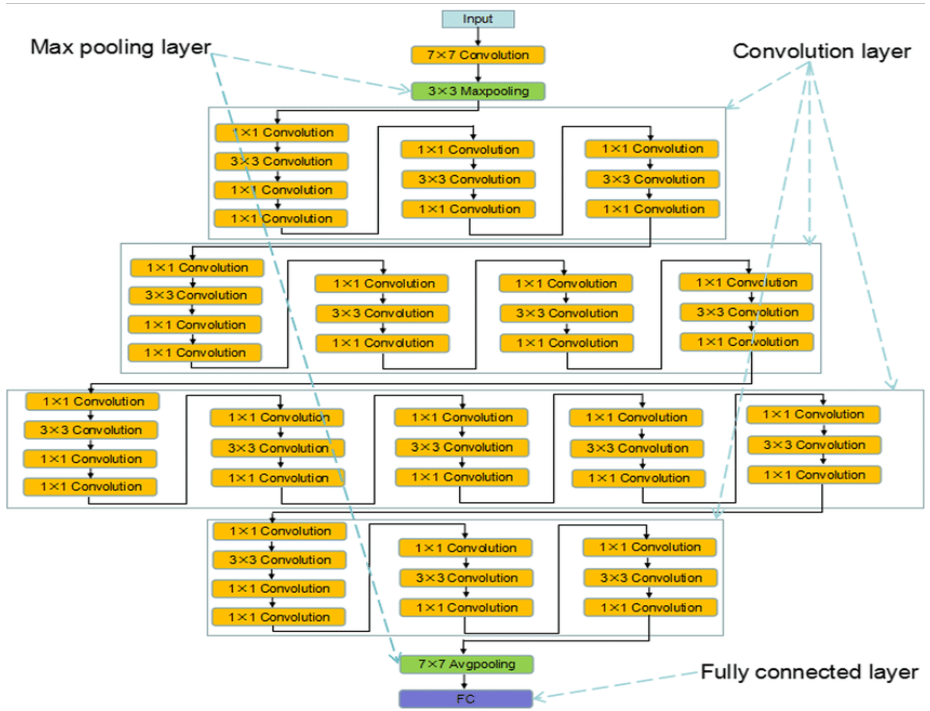
- 0.1% points for ResNet50 for a total of 78.7%
- 0.8% points for ResNet200 for a total of 81.4%

Research Question and Scope

- Is it possible to obtain in-domain like accuracies, with the Ingwersen model used as threshold for the in-domain model, for the general-domain model, using only a limited in-domain dataset?

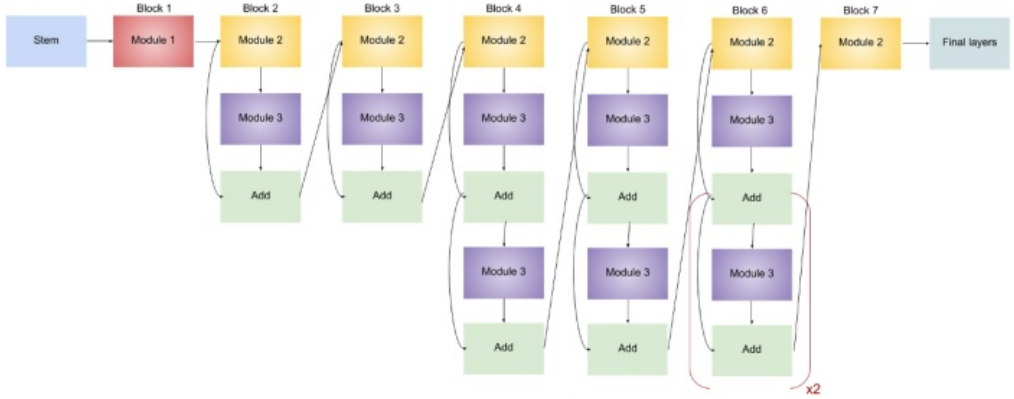
ResNet50 Model:

24 million total parameters



EffeciencNetB0 Model:

4 million total parameters



Methodology: General-Domain ImageNet

The general-domain model has been pre-trained on the ImageNet Dataset

ImageNet dataset consist of various different objects, and consists of 14 million labeled images, spread across 20000 categories.

The ResNet50 model recognizes 1000 classes from this dataset.

(a) ImageNet Synset: One sample image from each category

(b) Corel-1000 Dataset: Sample images from each category

(c) Caltech-256 Dataset: One sample image from each category

(d) Caltech-101 Dataset: One sample image from each category

Methodology: In-Domain Places365

The VGG-16 model used in Ingwersen were pre-trained on the in-domain Places365 dataset.

Places365 consists of 18 million images.

The Places365 dataset is spread across 385 different scene categories.

The VGG-16 Places365 model already recognizes various kinds of aesthetics in rooms.

Methodology: In-Domain Dataset Danish Real Estate (DRE)

The limited in-domain DRE dataset consists of 21000 images. Including 6500 labaled images, spread across the 6 different classes.

Images were harvested in 2019 by Ingwersen.

The dataset were labeled by Ingwersen, and the author PFJ.

The in-domain dataset should be used to teach a CNN to identify similarities in rooms, to improve customer engagements for real estate agencies.

(a) Label: Kitchen **(b)** Label: Living Room **(c)** Label: Dining Room

(d) Label: Bathroom **(e)** Label: Bedroom **(f)** Label: Entre

Methodology: Measuring Accuracies for Similarity

The Cosine Similarity measurement is usually used for naïve recomender systems, but proved to be useful when measuring similarities based on the feature vectors extracted by the proposed models.

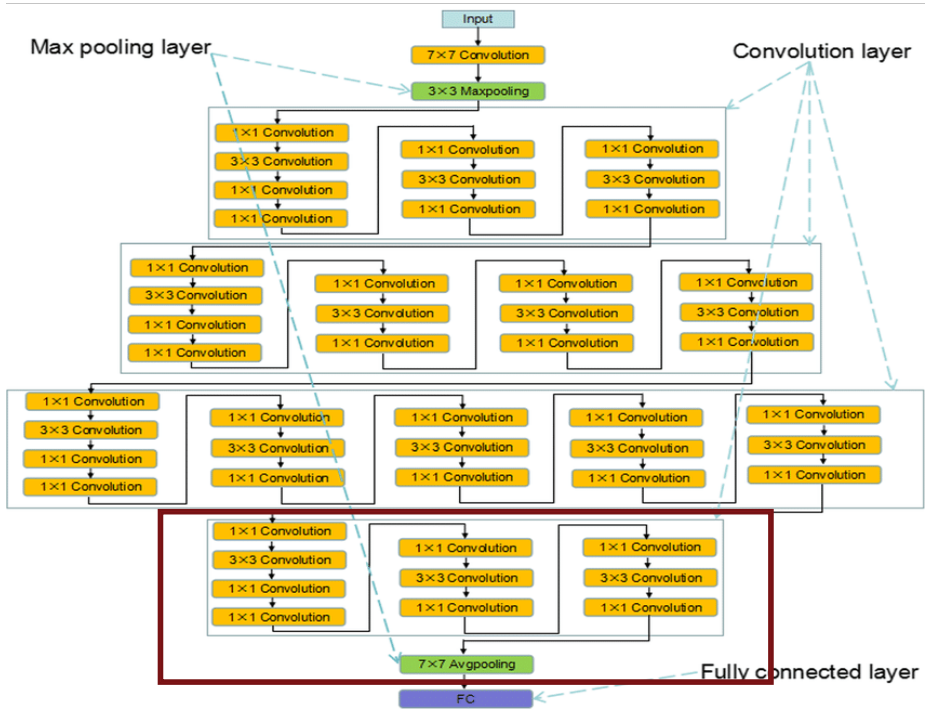
$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Methodology: Transfer Learning

By using transfer learning, the last complex layers are trained to learn meaningful representations of the images.
- Red boxes ilustrate the trained part of the models.

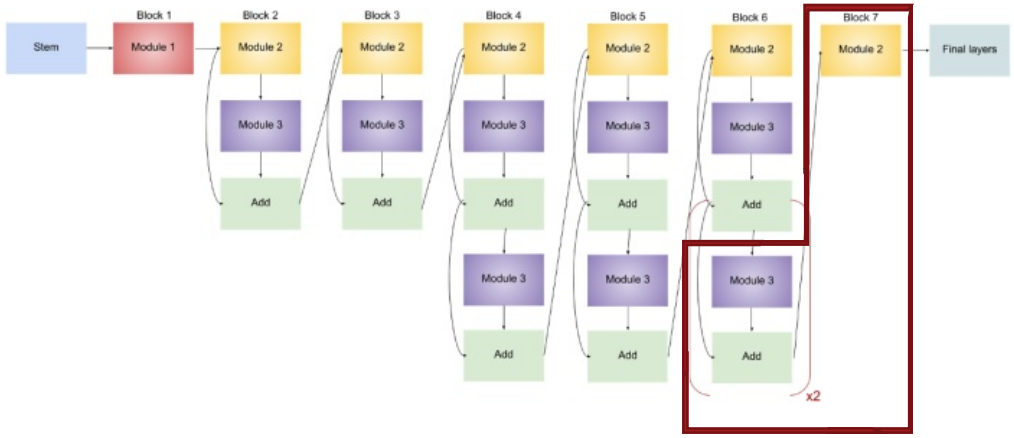
ResNet50 Model:

15 million trained parameters



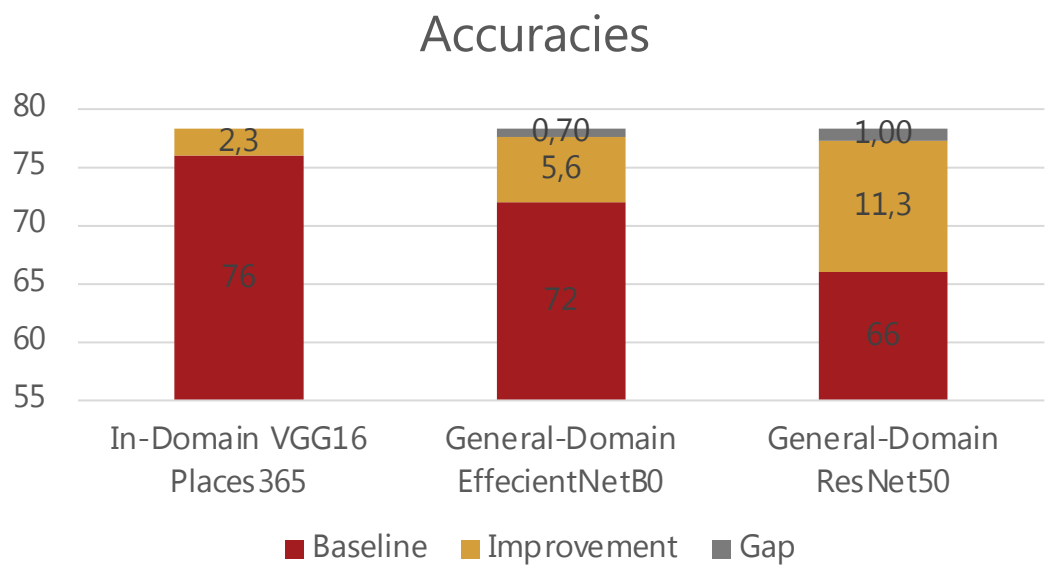
EffeciencNetB0 Model:

2 million trained parameters



Findings: Accuracies

The top-1 accuracy for the general-domain model achieved:



The size of the feature vector produced by the models had a significant saying in this.

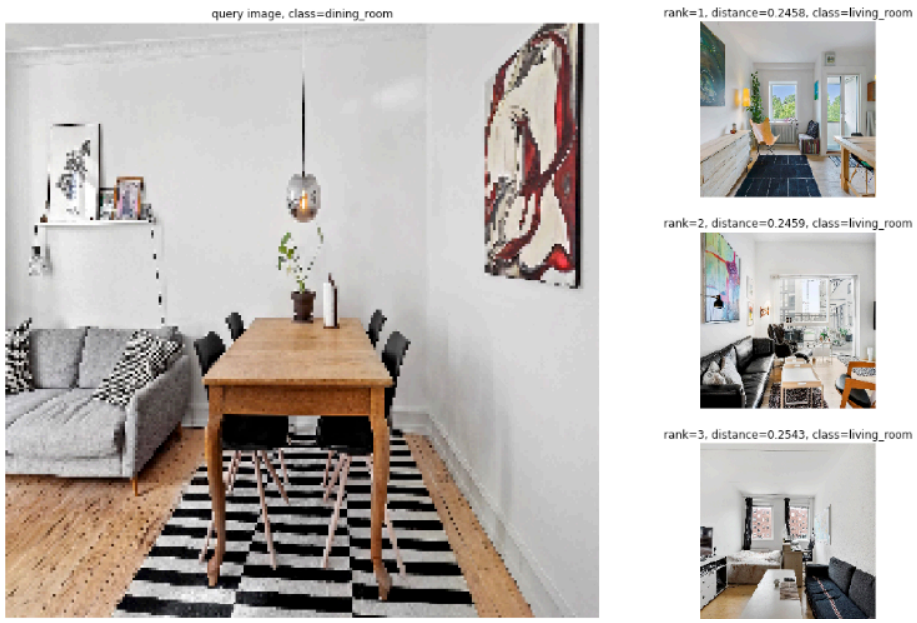
Both in Ingwersen and in this study it is suggested that the smaller the feature vector produced the higher the top-1 accuracy, scaling down to a certaint size.

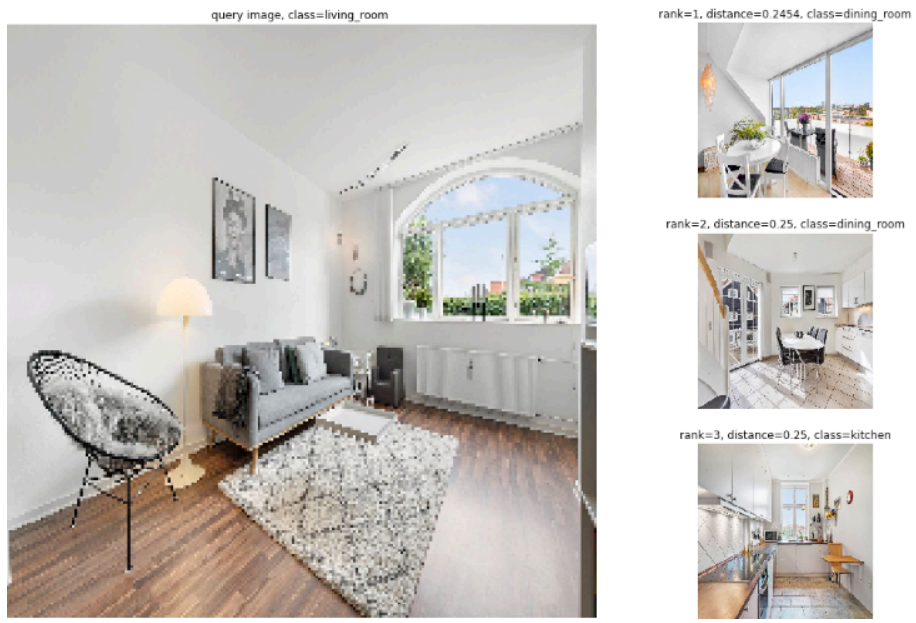
- In-Domain VGG16 Places365, size 512
- General-Domain EffecientNetB0, size 1280
- General-Domain ResNet50, size 2048

Findings: Visualisation of Untrained model (ResNet50)

The untrained general-domain ResNet50 model has not been able to capture meaningful representations of room.

The closest rooms does not have any noticeable similarities, and also seems to be predicting wrong types.

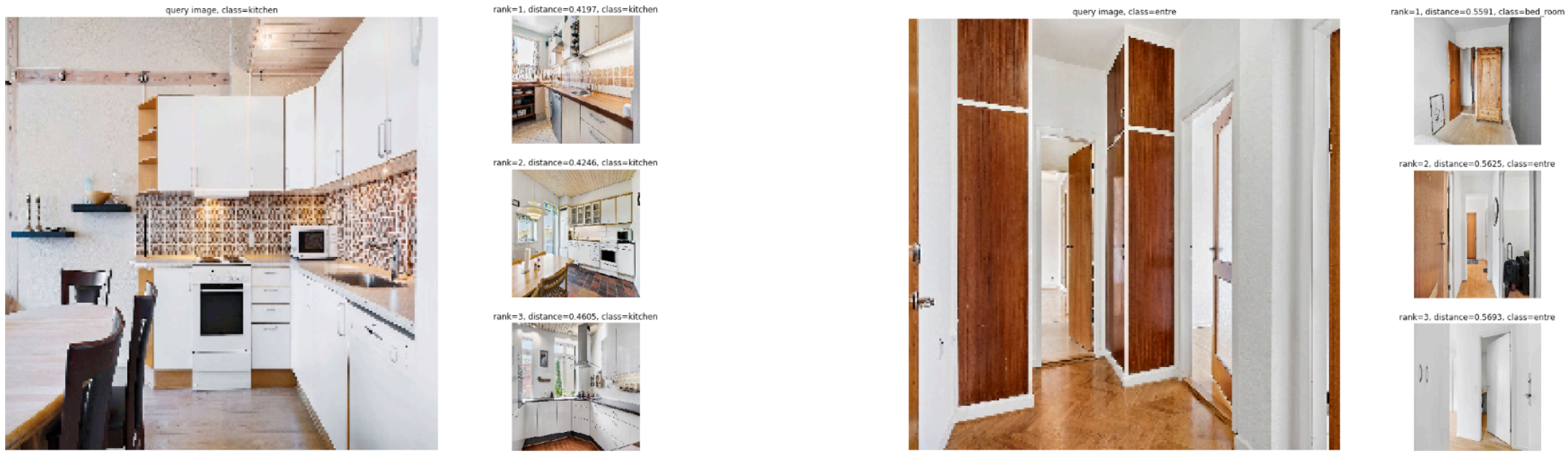




Findings: Visualisation of Trained EffecientNet

This model proposed in this study would successfully understand the aesthetics of the different rooms.

The model recognizes key features in rooms, including tiles in kitchens, and recognizes wooden panels, and wooden floor.



Findings: Visualisation of Trained ResNet50

The model proposed in this study would successfully understand the aesthetics of the different rooms.

The model recognizes key features in rooms, including tiles in the bathroom, and recognizes the spaciousness of bedrooms.

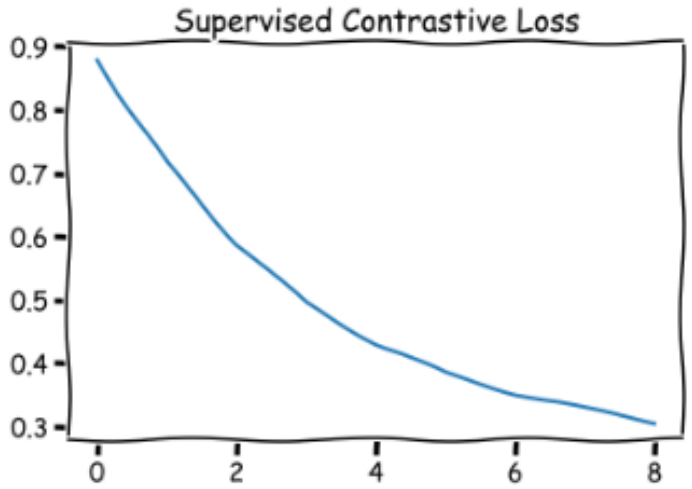
Limitations

Khosla Supervised Contrastive Learning

The model did finish training, however the top-1 accuracy for the model only achieved a top-1 accuracy of:

- 54.2%

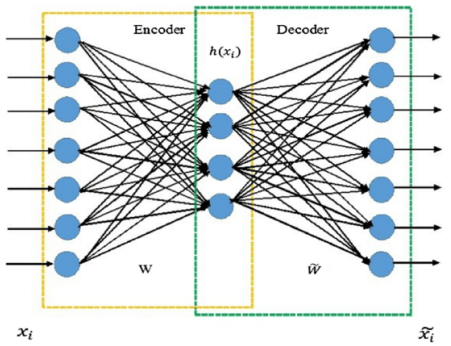
The contrastive loss might work against the pre-trained weights.



Ingwersen

The approach used to train the models in Ingwersen were not according to normal transfer learning practise.

- The entire network were trained
- The preprocessing schemes were different from the ones used to pretrain the network
- The autoencoder were trained using outdated feature vectors after first iteration.



Conclusion

- It was possible to train a general-domain model to obtain in-domain like accuracies for similarity measurements.
- The model proposed could help real estate agencies to increase customer engagement through a similar implementation, using only limited data sizes.
- Further investigation of unsupervised training would help real estate agencies to scale the model and reduce the neccesary manual labour needed when using supervised training.
- The model proposed in Ingwersen might be possible to be improved, by using proper practise for transfer learning and autoencoders.
  - Only training last layers of the CNN, as training too much proved to decrease the accuracy in this study.
  - Fix bug introduced with autoencoders, by updating feature vectors after each batch iteration.

Questions