Thesis Title

sub-title

MARCUS KLASSON

Doctoral Thesis
Stockholm, Sweden, 2020

KTH Royal Institute of Technology
School of Electrical Engineering and Computer Science
Division of Fusion Plasma Physics
SE-10044 Stockholm
Sweden

TRITA-EECS-AVL-2020:4
ISBN 100-

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av Teknologie doktorexamen i elektroteknik fredagen den 18 januari 2020 klockan 14.00 i Sal F3, Lindstedtsvägen 26, Kungliga Tekniska Högskolan, Stockholm.

Tryck: Universitetsservice US AB

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords: Lorem, Ipsum, Dolor, Sit, Amet

Sammanfattning

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

hej

List of Papers

A ***A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels***
Marcus Klasson, Cheng Zhang, Hedvig Kjellström
In *IEEE Winter Conference on Applications of Computer Vision (2019)*

B ***Using Variational Multi-view Learning for Classification of Grocery Items***
Marcus Klasson, Cheng Zhang, Hedvig Kjellström
In *Patterns, Volume 1(8) (2020)*

C ***Learn the Time to Learn: Replay Scheduling for Continual Learning***
Marcus Klasson, Hedvig Kjellström, Cheng Zhang
Under submission

D ***Meta Policy Learning for Replay Scheduling in Continual Learning***
Marcus Klasson, Hedvig Kjellström, Cheng Zhang
Under preparation

Acknowledgements

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Acronyms

List of commonly used acronyms:

AE	Acronym examples
CL	Continual Learning
CNN	Convolutional Neural Network
RL	Reinforcement Learning
VAE	Variational Autoencoder

Contents

Part I

Overview

Chapter 1

Introduction

Chapter 2

Background

The goal with this thesis is to provide machine learning methods for recognizing objects from images. Machine learning is a field within Artificial Intelligence where computer programs learns from experiences how to make predictions in new situations. There are three essential parts to enable the computer to make predictions with machine learning. Firstly, we need data representing the scenarios where we have objects that we wish to predict what they are. Secondly, we need a model that learns how to make the decisions based on the provided data. Thirdly, we need a learning algorithm for fitting the model to the data we have such that good and sensible decisions can be made on future data. Machine learning has proven to be successful on various types of data, including, images and video, text, and audio, and there exists many different kinds of models and algorithms for learning decision-making from data.

One of the main goals with machine learning is to have models that generalize to unseen data and events. However, there are several challenges that has to be tackled to achieve this goal. The first challenge is to obtain datasets that represent the events that the model should make predictions for. Machine learning models often require vast amounts of examples to learn from, and also, the examples should be annotated with some information describing each example in order to ease the learning. But even if we have large datasets, we must have models that have the capacity of preserving the knowledge gained from the dataset. Furthermore, we must have algorithms that can train the model from the huge amount of data in computationally efficient both time- and processing-wise. Especially for visual data, it has become much cheaper to obtain vast amounts of images and videos from the internet. Occasionally, these can be annotated through search words or, alternatively, from crowdsourcing. Moreover, computational power has also become cheaper through smaller and more efficient micro-processors, semi-conductors, and cloud computing. Deep learning [1] is a class of machine learning models based on neural networks that are capable of learning from large and high-dimensional datasets due to their capacity. They

are trained using an optimization algorithm called Stochastic Gradient Descent (SGD) [2] which works well for large-scale data and can be applied on graphical processing units (GPUs) with recent machine learning programming libraries, such as TensorFlow, PyTorch, and Jax. However, deep learning still faces lots of challenges in generalization, especially when they are applied in environments that were not present in the training data, and it is still an open research problem on how to make them generalize better.

There are several approaches for enabling better generalization for deep learning models. A good start is to collect datasets that are similar and represent the events in the environment where the model will be deployed. Related to this, one can also collect different data types from various modalities, such as visual signals in the form of images and video as well as natural language which can be written or spoken, if these are available in the data collection process and are sensible for the task to be solved. Multimodal machine learning opens up the possibility of learning correspondences between the different data types to gain better understanding of the phenomenon of interest [3, 4], which can help the model to be more accurate and robust. However, in order to enhance the utility of machine learning models, they should be capable of continuously updating their knowledge as many environments where object recognition is useful are ever-changing [5, 6]. We should build models that can add new objects of interest to recognize as well as delete concepts that are obsolete or non-relevant. It would also be useful if we could update models with personalized objects to recognize to narrow down the scope of items to recognize for object recognizers to make the tasks easier.

In this chapter, we cover related works on datasets on object recognition both with iamge and text data in Section XX. Next, we provide a description of machine learning, especially deep learning, models that were used in the included papers in Section XX. In Section XX, we discuss the setting of continual learning for updating the knowledge of machine learning models that is aiming to make the models capable of handling ever-changing environments as a step towards to enabling life-long learning.

2.1 Datasets for Object Recognition

The increasing accessibility of large-scale image datasets have enhanced the possibility for applying machine learning in various applications for visual recognition. One of the most well-known datasets in computer vision the is Imagenet database [7] introduced in 2009 which is still used for benchmarking models on large-scale image classification. Imagenet was collected through image search results from the Internet and then further assessed by crowdsourcers from Amazon Mechanical Turk to achieve higher quality of the collected images as well as labeling them. There has been several efforts to create more large-scale vision datasets, e.g., Pascal VOC [8], Microsoft COCO [9], Visual genome [10], which in addition to object classes include object attributes, bounding boxes for detecting

objects, and pixel-wise segmentation masks. Additionally, there exists datasets where images have been combined with text descriptions of things that are present and where they are located in relation to each other for image captioning and visual question answering tasks, e.g., Flickr30k [11], VQA [12], GQA [13], and Microsoft COCO Captions [14]. These publicly accessible datasets are one of the major reasons for enabling machine learning and computer vision research at a large scale which opens up for developing products that can be deployed on everyday products, such as mobile phones.

In order to deploy machine learning models in real-world scenarios, we first need training data that are close to the occurring events specific for the application where the model will be used. But even if we can train a model on huge amounts of images, it is extremely challenging to provide examples that cover all possible scenario that can happen or every different shape and color an object can have. Furthermore, other circumstances such as lighting conditions, occlusions, and other objects in the surroundings of objects of interest which can be hard to control can also make the recognition performance less accurate. This is critical when training models where the user is strongly relying on the recognition performance, such as in assistive vision systems for blind or low-vision people. As many of the popular image datasets for computer vision contains images from the Internet, there can be a large gap between the training images and the images that will be seen after deployment as Internet images often have good conditions and the objects are fully visible and centered. However, when the model would be used in the real-world, the objects could be occluded or not fully visible in the image and be disturbed by objects in the background. Therefore, it can be critical for the performance and robustness of the model to train it on images that are tailored for the scenarios where it will be applied.

There have been several attempts to build datasets with images collected by visually impaired to obtain more realistic datasets with potential scenarios. VizWiz [15] is one of the first large-scale image datasets with mobile phone images taken by blind people where the user have asked a question about each image that are answered by crowdworkers. This dataset is very challenging as questions asked by the collectors can be unanswerable since the objects can be occluded or even out of frame. The ORBIT dataset [16] is a more recent dataset that have built a dataset of videos from blind or low-vision people of their personal items, e.g., keys, wallets, remote controls, to enable few-shot learning of personalizable object recognizers. This dataset is the first to contain video recordings which potentially can be more user friendly as this allows the user to rotate the items that could yield more accurate performance.

2.2 Machine Learning

In this section, we give a brief introduction to basics in machine learning and some mathematical notation that will be used in this chapter. The overall goal is

to learn some phenomena from a set of N data points $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ where $\boldsymbol{x}_i \in \mathbb{R}^D$ has D dimensions for $1 \leq i \leq N$. The learning part is referred to as the training phase where we the goal is to fit the machine learning model to the data points, or observations, X. The objective of the training phase is determined by what the target task is tasks that we wish to solve using the model. Next, we give a brief description of three paradigms in machine learning with different end goals depending on what we wish to explore.

Supervised Learning. Many types of data have a corresponding target $\boldsymbol{y}$ that explains the content of each data point. Cases where the target comes from a discrete number of classes and the goal is to determine which class that the data point corresponds to are called classification problem. An classic example of a classification problem is distinguishing whether there is a cat or dog in images. If the target $\boldsymbol{y}$ is a continuous valued variable and the goal is to predict this value, then we have a regression problem, where an example can be predicting the outdoor temperature tomorrow given the current temperature. The goal for both of these problems is to learn a function $f(\boldsymbol{x})$ that can classify or predict the given target data as accurately as possible.

Unsupervised Learning. For some problems, we may only have the data points for our disposal where we are interested in finding some patterns in the given data. The goal in unsupervised learning problems can be to discover similar groups with clustering techniques, or estimating the distribution of the data known as density estimation. The practitioner typically needs to define some assumptions about the data, e.g., how the similarity between two data points should be measured, before we can execute the algorithm for discovering the patterns.

Reinforcement Learning. In this paradigm, we are concerned with decision-making where the goal is to take actions that maximize some reward. The decision-making is modeled by a policy which bases the action selection on observations that are collected by interacting with the task environment through the selected actions. One main challenge is how to handle the trade-off between exploration of different actions in new situations and exploitation by selecting actions where the agent already has experienced good reward signals. Furthermore, the reward signal can be received either in dense or sparse forms, where sparse rewards are typically more challenging to learn from and are less sample-efficient.

MK: Could be nice with some intro to data distributions some where, just like explaining what $x \sim p(x)$ for example and explaining that x can be any image in the world or in some environment of the world.

2.3 Deep Learning

Neural networks is a class of machine learning models which popularity have grown immensely due to their ability to learn from large and high-dimensional datasets. Moreover, neural networks have been successfully applied in various number of fields in computer vision [17,18], natural language processing [19], and reinforcement learning [20, 21]. These models are constructed by stacking layers of parameters that extract representations of the input data until reaching the final layer that outputs the target answer from the queried input. For example, the operation for passing the input data $\boldsymbol{x}$ through the first layer can be denoted as the matrix multiplication $\boldsymbol{h} = \boldsymbol{W}_1\boldsymbol{x}$ where $\boldsymbol{W}_1$ are the weights of the first layer. An essential part for enabling neural networks to learn more interesting non-linear functions is to add an activation function right after the matrix multiplication of each layer, otherwise the neural network would only be capable of learning linear functions. A common activation function is the $a(\boldsymbol{x}) = \max(0, \boldsymbol{x})$, or the so called Rectified Linear Unit (ReLU) activation, which outputs $\boldsymbol{x}$ when $\boldsymbol{x} > 0$ or otherwise zero. Stacking two layers together in a neural network with such activation between then becomes $\boldsymbol{h} = \boldsymbol{W}_2 \max(0, \boldsymbol{W}_1\boldsymbol{x})$. In classification problems, the output layer outputs the score for which class the input $\boldsymbol{x}$ could belong to. The weight parameters $\boldsymbol{W}_1, \boldsymbol{W}_2$ are learned with SGD, and their gradients are derived using the chain rule and computed using backpropagation (see [1] for an introduction to backpropagation).

MK: Feel like I would like to include some brief info on CNNs and RNNs, but just a paragraph long for each. Maybe also describe the Autoencoders a bit.

Variational Autoencoders

Generative models are used for approximating a data distribution p_{data} from a given set of samples $\mathcal{D}$. In most cases, we parameterize the distribution with $\boldsymbol{\theta}$ and learn the parameters from the given data by minimizing some distance metric between the estimated distribution $p_{\boldsymbol{\theta}}$ and p_{data}. In deep generative models, the distribution p_θ is parameterized by a neural network where the parameters $\boldsymbol{\theta}$ represent the weights and biases in the network. These models can be broadly divided into three classes of models, namely Variational Autoencoders (VAEs) [22], Generative Adversarial Networks (GANs) [23], and Normalizing Flows [24]. Their commonality is that they are based on latent variable models where it is assumed that the observed data is generated from some hidden process from a simpler distribution than p_{data}. However, which class of models to select depends on the application and the goals with the task. In this thesis, we focus on VAEs because of their capability of learning lower-dimensional representations.

A key ingredient in learning generative models is to introduce a latent variable $\boldsymbol{z}$ where we assume that $\boldsymbol{z}$ is related to the observed variable $\boldsymbol{x}$ through the data generation process. We can still estimate the parameterized distribution when

introducing the latent variables since

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\, d\boldsymbol{z} = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z})\, d\boldsymbol{z}, \tag{2.1}$$

where we incorporate $\boldsymbol{z}$ to obtain the joint distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ and use the chain rule for probabilities to obtain the likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ and the prior distribution $p(\boldsymbol{z})$. The prior $p(\boldsymbol{z})$ is where we can define our assumption of how the underlying hidden reasons are distributed by selecting a simple and well-known distribution for this space, e.g., a Gaussian distribution. The goal with latent variable models is often to compute the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ over the latent variables given the data $\boldsymbol{x}$ which can be done using Bayes' rule

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}. \tag{2.2}$$

Unfortunately, the evidence $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is very hard to compute as it requires calculating the integral $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\, d\boldsymbol{z}$ over all dimensions of the latent variable space. To overcome this issue, we will propose a distribution $q_{\boldsymbol{z}}$ that should approximate the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$. Variational Inference (VI) [25, 26] is a method for approximating probability densities through optimization which is used in VAEs for approximating the posterior. Next, we give a description of how the posterior is approximated.

The learning objective in VAEs is based on the derivation of the marginal density, or evidence, in Equation 2.1. As the evidence is intractable to compute exactly, we will instead maximize a lower bound on the evidence that is called the Evidence Lower BOund (ELBO). The ELBO is derived in log-space such that we can apply Jensen's inequality to obtain the lower bound when we have introduced the approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ parameterized by $\boldsymbol{\phi}$. We begin be deriving a general expression for the ELBO:

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\, d\boldsymbol{z} \\
&= \log \int \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\, d\boldsymbol{z} \\
&= \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right] \\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right],
\end{aligned} \tag{2.3}$$

where we applied Jensen's inequality between the third and fourth line to obtain the ELBO. This expression can be derived further to obtain the common VAE

objective

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - D_{\mathrm{KL}}[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})], \end{aligned} \tag{2.4}$$

where the second term is the Kullback-Leibler (KL) divergence between the approximate posterior and the prior. The KL divergence can be computed exactly when selecting both $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $p(\boldsymbol{z})$ to be Gaussian densities, while the expectation of the log-likelihood can be estimated with Monte Carlo approximation. Furthermore, the reparameterization trick [22, 27] is used to enable computing the gradients through the sampling step when estimating the log-likelihood that can be used for backpropagation.

Multi-view Learning

2.4 Continual Learning in Neural Networks

Chapter 3

Summary of Included Papers

In this chapter, we provide a summaries of the included paper for this thesis. Paper A and B are connected through the Grocery Store dataset where we present the work and then perform an ablation study over which modalities in the dataset that are useful for training classifiers. In Paper C and D, we focus on continual learning (CL) and present a new setting that aims to fill the gap between CL research and real-world problems as well as a method for doing so.

A A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels

Authors: Marcus Klasson, Cheng Zhang, Hedvig Kjellström.

Summary. We collect a dataset with natural images of raw and refrigerated grocery items taken in grocery stores in Stockholm, Sweden, for evaluating image classification models on a challenging real-world scenario. The data collection was performed by taking photos of groceries with a mobile phone to simulate a scenario of grocery shopping using an assistive vision app. Furthermore, we downloaded iconic images and text descriptions of each grocery item by web-scraping a grocery store website to enhance the dataset with information describing the semantics of each individual item. the items are grouped based on their type, e.g., apple, juice, etc., to provide the dataset with a hierarchical labeling structure.

We provide benchmark results evaluated using pre-trained and fine-tuned CNNs for image classification. Moreover, we take an initial step towards utilizing the rich product information in the dataset by training the classifiers with representations where both natural and iconic images have been combined through a multi-view VAE.

Author Contributions. CZ and HK presented the idea and the data collection procedure for the natural images and web-scraped information. MK performed

the data collection including visiting the grocery stores for taking the natural images and the web-scraping of the grocery store website for iconic images and text descriptions. MK performed all the experiments. All authors contributed to discussing the results and contributed to writing the manuscript.

B Using Variational Multi-view Learning for Classification of Grocery Items

Authors: Marcus Klasson, Cheng Zhang, Hedvig Kjellström.

C Learn the Time to Learn: Replay Scheduling for Continual Learning

Authors: Marcus Klasson, Hedvig Kjellström, Cheng Zhang.

D Meta Policy Learning for Replay Scheduling in Continual Learning

Authors: Marcus Klasson, Hedvig Kjellström, Cheng Zhang.

Summary.

Author Contributions. CZ presented the idea.

Chapter 4

Discussion and Conclusions

4.1 Conclusions

4.2 Future Work

- Video data for object recognition instead of images for making systems easier to use
- Federated Learning for decentralizing model updates
- Uncertainty Quantification - How to make the classifiers trustworthy?

Bibliography

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[2] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186, Springer, 2010.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[4] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[5] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[7] 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*, pp. 248–255, Ieee, 2009.

[8] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting

language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[11] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[12] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[13] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

[14] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[15] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

[16] D. Massiceti, L. Zintgraf, J. Bronskill, L. Theodorou, M. T. Harris, E. Cutrell, C. Morrison, K. Hofmann, and S. Stumpf, "Orbit: A real-world few-shot dataset for teachable object recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10818–10828, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[21] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[24] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*, pp. 1530–1538, PMLR, 2015.

[25] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[26] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[27] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*, pp. 1278–1286, PMLR, 2014.

Part II

Included Papers

Paper A

Name for Paper A

Marcus Klasson, Cheng Zhang, Hedvig Kjellström

Abstract

Abstract aby stract

A.1 Introduction

hej hej här är en artikel

what do you think this is?

hello hello city

Paper B

Name for Paper B

Marcus Klasson, Cheng Zhang, Hedvig Kjellström

Abstract

Abstract aby stract

B.1 Introduction

hej hej här är en artikel

what do you think this is?