



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 Technologie doktorexamen i elektroteknik fredagen den 18 januari 2020 klockan 14.00 i Sal F3, Lindstedtsvägen 26, Kungliga Tekniska Högskolan, Stockholm.

© Marcus Klasson, date

Tryck: Universitetsservice US AB

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur 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, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur 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, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur 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

List of Papers	iii
Acknowledgements	v
Acronyms	vii
Contents	1
 I Overview	 3
1 Introduction	5
1.1 Vision Impairments	5
1.2 Object Recognition for Assistive Vision	5
1.3 Thesis Contributions	5
1.4 Thesis Outline	5
 2 Background	 7
2.1 Datasets for Object Recognition	8
2.2 Modeling	10
 3 Summary of Included Papers	 11
A A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels	11
B Using Variational Multi-view Learning for Classification of Grocery Items	12
C Learn the Time to Learn: Replay Scheduling for Continual Learning	12
D Meta Policy Learning for Replay Scheduling in Continual Learning .	12
 4 Discussion and Conclusions	 13
4.1 Conclusions	13
4.2 Future Work	13
 Bibliography	 15

II Included Papers	17
A Name for Paper A	A1
A.1 Introduction	A1
B Name for Paper B	B3
B.1 Introduction	B3

Part I

Overview

Chapter 1

Introduction

1.1 Vision Impairments

1.2 Object Recognition for Assistive Vision

1.3 Thesis Contributions

1.4 Thesis Outline

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 learn 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 have 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 image 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 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 is the 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 Modeling

Deep Learning

Multi-view Learning

Continual Learning

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.

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

A2

PAPER A. NAME FOR PAPER A

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

B4

PAPER B. NAME FOR PAPER B

what do you think this is?