C. Käding, E. Rodner, A. Freytag, J. Denzler
Watch, Ask, Learn, and Improve: a lifelong learning cycle for visual recognition. ESANN 2016
The final publication is available at the ESANN webpage.

Watch, Ask, Learn, and Improve: a lifelong learning cycle for visual recognition

Christoph Käding, Erik Rodner, Alexander Freytag, Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Germany

Abstract. We present WALI, a prototypical system that learns object categories over time by continuously watching online videos. WALI actively asks questions to a human annotator about the visual content of observed video frames. Thereby, WALI is able to receive information about new categories and to simultaneously improve its generalization abilities. The functionality of WALI is driven by scalable active learning, efficient incremental learning, as well as state-of-the-art visual descriptors. In our experiments, we show qualitative and quantitative statistics about WALI's learning process. WALI runs continuously and regularly asks questions.

1 Introduction

In recent years, we observed a gigantic leap in machine learning and computer vision due to the success of deep learning. The implicit combination of feature learning and model training lead way to a series of applications. Impressive examples are large-scale image classification [1] and pixel-wise classification of images [2]. Furthermore, challenging application areas such as medical data analysis [3] or earth observations [4] benefit from recent advances in machine learning and computer vision.

As impressive as these results are in their respective fields, our current technology still suffers from two crucial drawbacks. First of all, we require massive amounts of labeled data to train deep networks [1] in a supervised manner. In direct consequence, training is computationally expensive even with dedicated hardware using latest GPU generations. Furthermore, our current solutions follow the closed-world assumption [5]. As an example, we often assume each and every possible object category to be contained in ImageNet [6]. However, our world is changing in almost any possible way. Thus, data distributions are not fixed and new categories appear over time or existing categories vary in their visual appearance (*e.g.,* the latest Mercedes is not yet in ImageNet).

In this article, we present a system that overcomes both drawbacks. Particularly, we draw inspiration from the way children acquire knowledge over time. Similar to this astonishing ability, we argue for the necessity of lifelong learning systems which curiously adapt to new scenarios. In summary, our proposed prototype consists of four steps to simulate human learning: (1) Watch, (2) Ask, (3) Learn, and (4) Improve – WALI.

Our main motivation comes from observing children interacting with their environment. During day-time, they are continuously facing new impressions. Additionally, children obtain a variety of additional guiding information, *e.g.,* in terms of direct supervision ("That's an elephant!") or negative supervision ("No, that's not a giraffe!").

This research was supported by grant DE 735/10-1 of the German Research Foundation (DFG)

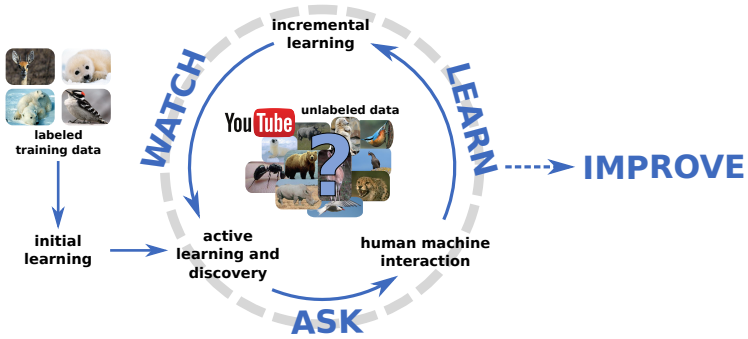

Fig. 1: Overview of the lifelong learning cycle used by WALI.

Thereby, children accumulate a variety of new "labeled data", which is ultimately reorganized during their sleep phases.

In this paper, we build on the previous observations to design lifelong learning systems. Particularly, we develop a system faced with continuous streams of `youtube` video data (Watch). We then focus on a specific form of guiding information. Similar to child-parent-interactions, we allow the system to actively select few examples for manual annotation (Ask). Obtained feedback is used to update the current model (Learn). Ultimately, this never-ending cycle allows for adapting to new scenarios (Improve). A visualization is given in Fig. 1

Related work A highly related lifelong learning system is NEIL [7], which automatically mines the web for new visual concepts and uses image search engines to associate images with weak labels. In contrast, we focus on an interactive system which asks a human annotator without relying on text data. Another related system, called VisKe [8], gets online suggestions for things it should discover. It is not restricted to relationships between nouns and instead tries to verify verb-based relationships utilizing object detectors and factorization graphs. Although WALI also incorporates human input, the provided feedback is used to label images and not to guide the system directly during its discovery. In contrast to more general curiosity driven system like [9], our system focuses currently on learning image classifiers incrementally from human feedback.

2 Watch, Ask, Learn, and Improve

The blueprint for WALI leaves several choices for realizing each involved aspect. In the following, we introduce specific solutions and explain our design choices.

Watch: feeding WALI with a continuous stream of data WALI requires a continuous stream of unlabeled images for lifelong learning. We obtain this data stream by automatically downloading videos from `youtube` using the official API. Since we currently focus on animal categorization in our experiments, we let WALI search for "animal documentary". To reduce the amount of redundant data, we sample images from each video by taking every 10^{th} frame.

All images obtained during the recent watch session are available for selection. In addition, also data from previous watch sessions might still contain valuable information. To access this large pool of data, we equip WALI with a forgetting strategy. Therefore, a random set of 500 images which occurred in the last 5 hours is added to the pool of possible questions. Thereby, WALI has the possibility to still ask questions about interesting things seen in the past if recent recordings are insufficient. Besides practical advantages, we believe that this forgetting-aspect is related to human learning.

As visual descriptors, we make use of state-of-the-art features obtained from activations of pre-trained convolutional neural networks. For various tasks, CNN activations have been proven to deliver expressive image descriptions [6]. By using those activations, we are able to describe even unknown classes adequately [10]. In particular, we use the open-source `caffe` framework [11] along with the AlexNet network originally trained with images from the ImageNet challenge dataset [6]. Extracting relu7 activations from the entire image and normalizing them yields $4{,}096$-dimensional image representations.

Ask: feedback with limited supervision The key feature of WALI is to actively select images for labeling by human annotators. While this opportunity enables the incorporation of expert feedback, it comes with several challenges. First of all, a random selection of images likely results in redundant questions. Furthermore, examples of unknown categories should be identified for selection to expand the current knowledge. Finally, not all questions can be answered by human annotators [12]: either due to their limited knowledge (unknown) or if selected images do not contain a valid category (non-categorial). In consequence, selecting images which likely lead to an information gain is far from trivial.

To solve these issues, a variety of active learning techniques have been introduced over the last decades (see [13] for an overview). In common active learning scenarios, instances from a large pool of unlabeled examples are evaluated whether they likely result in an increase of the classifier performance once they have been labeled and added to the training set. After selecting the best-ranked example for annotation, received information is used to update the classification model and to re-estimate active learning scores for subsequent iterations.

For WALI, we use linear one-vs-all classifiers $\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}$ for each class $k \in \mathcal{Y}_t^q$ which is known at time step t. Models are learned with class-balanced linear regression. Query images are selected according to the best vs. second-best strategy as proposed in [14] which scales even to large number of labeled and unlabeled examples. Thus, we evaluate the active learning score $q(\boldsymbol{x})$ for every unlabeled example $\boldsymbol{x}$ as follows:

$$\hat{k} = \operatorname{argmax}_{k \in \mathcal{Y}_t^q} \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x} \ , \tag{1}$$

$$q(\boldsymbol{x}) = \boldsymbol{w}_{\hat{k}}^{\mathrm{T}}\boldsymbol{x} - \operatorname{argmax}_{k \in \mathcal{Y}_t^q \setminus \{\hat{k}\}} \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x} \ . \tag{2}$$

The example with smallest score is selected for labeling which accounts for the intuitive preference of examples the current classifier is most uncertain about.

In contrast to traditional active learning scenarios, we additionally allow the annotator to reject examples for labeling. This is particularly important for lifelong learning and open-set recognition, since not every frame of a given video can be associated with

a semantic category. Furthermore, the human supervisor might not be interested in certain semantic categories.

Therefore, we follow [12] and add a rejection category comprised of all examples the annotator refused over time. To avoid frequent selection of examples which are likely to be rejected, we incorporate the probability that $\boldsymbol{x}$ belongs to the rejection class into the active learning criterion:

$$\tilde{q}(\boldsymbol{x}) = (1 - p(\text{rejection} \mid \boldsymbol{x})) \cdot q(\boldsymbol{x}) \ . \tag{3}$$

The probability is estimated using outputs of a binary classifier for the rejection class which is calibrated with the probit technique described in [12].

Learn: efficient and incremental For our lifelong learning scenario, we require classifiers which are efficient to evaluate and which easily scale to large numbers of training examples. Therefore, we use linear classifiers learned with linear regression. In particular, a quadratic loss function is used, which allows us to efficiently update the classifiers with new training examples without completely re-training it (see [15, Sect. 7–9] for details).

3 Experimental Results and Discussion

In the following, we show WALI's ability to quickly discover new classes and improve it's underlying classification model.

Quantitative analysis with ImageNet categories We compare three strategies for WALI: we either request labels from a human annotator after every 5 minutes for 10 actively selected instances (05/10-active) or we select randomly or actively 60 examples after 30 minutes (30/60-random and 30/60-active). Thus, all strategies result in the same number of posed questions.

For every image which has been selected by WALI, an annotator is allowed to specify the label which he estimates as most adequate. To still allow for quantitative analyses, we perform a manual matching of all provided labels to ImageNet synsets. Thereby, we can evaluate the accuracy of WALI's learned classifiers on all available ImageNet images of the corresponding synsets.

An important observation for lifelong learning is that the set of discoverable classes is unknown in advance. Let $\mathcal{Y}_t^q$ denote the set of classes which have been discovered by WALI at time step t with active selection strategy q. To allow for comparable evaluations, we merge all sets $\mathcal{Y}_t^q$ discovered in any experiment to test against: $\mathcal{Y}_\infty = \bigcup_{q,t} \mathcal{Y}_t^q$. As performance measure, we use the hierarchical ImageNet error [6] to consider confusions between semantically related classes. For evaluation, we show the number of discovered classes $|\mathcal{Y}_t^q|$ at each time step, the hierarchical error err-$\mathcal{Y}_t^q$ with respect to the currently discovered classes $\mathcal{Y}_t^q$, and the hierarchical error err-$\mathcal{Y}_\infty$ regarding all classes $\mathcal{Y}_\infty$. The results are given in Fig. 2 for several time steps in WALI's life.

First of all, we observe that err-$\mathcal{Y}_t^q$ increases over time. This is caused by an increasing number of known categories and resulting mis-classifications. Complementary, err-$\mathcal{Y}_\infty$ decreases with more accessible data as expected. We further observe that WALI gains less accuracy when we force a frequent selection from short videos (blue)

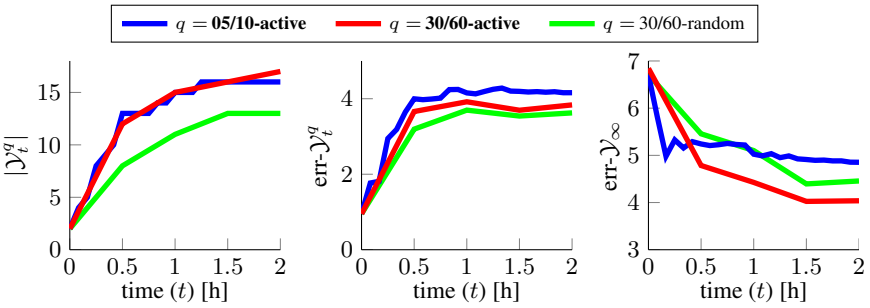

Fig. 2: Evaluation of WALI's performance on subsets of ImageNet. *Left*: number of discovered classes $|\mathcal{Y}_t^q|$. *Middle*: hierarchical error err-$\mathcal{Y}_t^q$ regarding known categories. *Right*: hierarchical error err-$\mathcal{Y}_\infty$ regarding all categories.

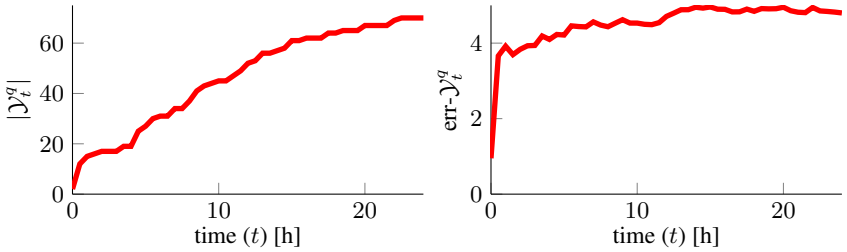

Fig. 3: Longterm evaluation of WALI's performance on subsets of ImageNet using the 30/60-active strategy. *Left*: number of discovered classes $|\mathcal{Y}_t^q|$. *Right*: hierarchical error err-$\mathcal{Y}_t^q$ with respect to known categories.

compared to longer clips (red). Furthermore, a mere random selection (green) leads to comparable classification accuracy but is clearly inferior regarding discovery of novel categories. Additionally, we performed a longterm experiment over 24 hours using the 30/60-active strategy which is shown in Fig. 3. We observe that WALI continuously discovers classes while err-$\mathcal{Y}_t^q$ is nearly constant.

Qualitative analysis We additionally show images in Fig. 4 for which WALI requested annotations during learning. As can be seen, WALI is able to successfully detect a multitude of novel categories over time.

4 Conclusions and Future Work

We showed the abilities of WALI, an existing lifelong learning system for visual recognition. WALI is able to actively acquire unlabeled video streams and to ask for relevant annotations. Currently, WALI is a prototypical implementation which runs continuously. Future work will focus on several extensions of WALI. Among others, we will integrate object localization, incremental training of convolutional neural networks, hierarchical classification, and richer annotations.

Fig. 4: Images requested by WALI for manual annotation. The current step is given in the top left corner. The category "noise" refers to cases, where the annotator refused labeling. Images were obtained from various `youtube` videos.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[3] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *arXiv preprint arXiv:1505.03540*, 2015.

[4] Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*, 2015.

[5] Heinrich Niemann. *Pattern analysis and understanding*. Springer Science & Business Media, 2013.

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[7] Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013.

[8] Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015.

[9] Varun Raj Kompella, Marijn Stollenga, Matthew Luciw, and Juergen Schmidhuber. Continual curiosity-driven skill acquisition from high-dimensional video inputs for humanoid robots. *Artificial Intelligence*, 2015.

[10] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, 2014.

[12] Christoph Käding, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *CVPR*, 2015.

[13] Burr Settles. *Curious machines: active learning with structured instances*. PhD thesis, 2008.

[14] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009.

[15] Ronald L. Plackett. Some theorems in least squares. *Biometrika*, 1950.