Efficient Deep CNNs for Cross-Modal Automated Computer Vision under Time and Space Constraints

Mohammadreza Amirian*†, Katharina Rombach*, Lukas Tuggener*‡, Frank-Peter Schilling*, and Thilo Stadelmann*
* ZHAW Datalab, Zurich University of Applied Sciences, Winterthur, Switzerland
† Institute of Neural Information Processing, Ulm University, Germany
‡ Università della Svizzera italiana, Lugano, Switzerland

Abstract

We present an automated computer vision architecture to handle video and image data using the same backbone networks. We show empirical results that lead us to adopt MOBILENETV2 as this backbone architecture. The paper demonstrates that neural architectures are transferable from images to videos through suitable preprocessing and temporal information fusion.
Keywords: AutoDL, automated deep learning, convolutional neural networks, MOBILENETV2, EFFICIENTNET

I. INTRODUCTION

By design of the evaluation metric, the recently completed AutoCV2 challenge aimed at finding *efficient* models for learning image and video datasets [1]. Consisting of five image datasets as well as three for video processing, it provided a range of multi-label and classification tasks such as object recognition, cancer-type classification, and motion detection. Goal was to develop a fully automated deep learning system able to learn any of the given tasks as quickly as possible. The training datasets were available for development and participants were allowed to make a limited number of submissions to the validation datasets. Final evaluations took place based on a set of hidden test datasets.

The Area under the Learning Curve (ALC) has been chosen by the challenge organizers as the evaluation metric; it is in favor of models which learn a given task quickly with low computational complexity. Therefore, the following directions of research seem most appropriate to improving the final performance:

- Light-weight models
- Sample efficiency for a quick training
- Preventing overfitting with low regularization
- Meta pretraining and generic architecture search

II. MODEL

We propose a modular, unified solution that tackles both image and video classification in order to deliver a quick and generic system for automated deep learning with different data types. While there is spatial information in each image, videos add temporal information in their sequence of frames. The backbone of the model is MOBILENETV2 [2] or EFFICIENTNET [4] for extracting spatial information from images or randomly selected consecutive frames from the video. This building block is the same for video and image classification and can thus be reused.

A. Image processing

To address variable image size, we first resize each image to the average size of the whole dataset. If the images have an edge longer than 128, they get resized such that the longer edge has length 128 while keeping the aspect ratio constant. Data augmentation is neglected based on empirical results since the gained accuracy cannot compensate for the slower training process in terms of ALC. The final proposed model uses the MOBILENETV2 as backbone for image processing followed by a one dimensional convolution, spatial average pooling and three fully connected layers for classification.

B. Video processing

An information fusion building block for videos processes the additional temporal information. Experiments were conducted with 3D convolutions (spatial and temporal information intertwined) as well as with 2D spatial convolutions followed by 1D temporal convolutions. It was further investigated at which depth these convolutions should be introduced to the network: on top of the final feature maps of the backbone model, or earlier. Instead of fixed pre-trained filters for the convolutions of the information fusion block, we propose an adaptive filter size and frame selection with respect to the

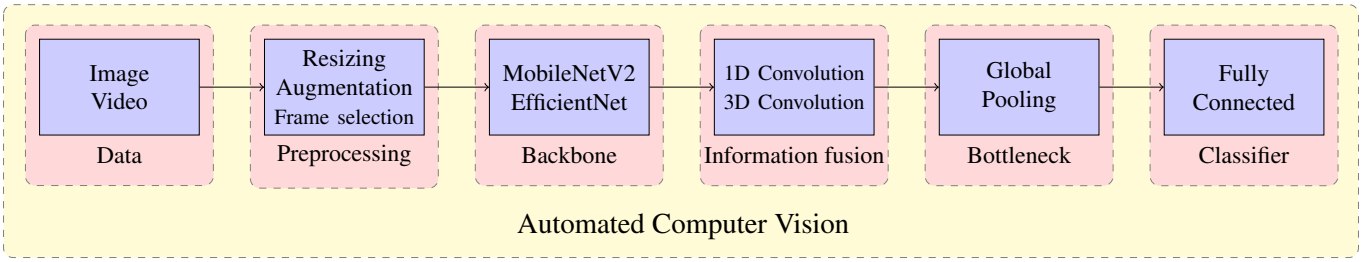

Fig. 1: Block diagram of the proposed automated computer vision framework.

Datatype		Image				Video		
Dataset		Chucky	Decal	Hammer	Pedro	Katze	Kraut	Kreatur
MOBILENETV2	Return Time	15.86s	20.35s	26.42s	25.53s	21.54s	21.29s	16.24
	ALC	**0.7536**	**0.7759**	**0.7562**	**0.7491**	**0.8508**	**0.6551**	**0.8678**
	NAUC	0.6853	**0.8568**	0.8538	0.8712	0.9487	**0.7433**	**0.9521**
	Overfit	Yes	Minor	No	No	No	No	No
EFFICIENTNET-MINI	Return Time	75.65s	76.53s	85.51s	83.48s	83.28s	82.77s	80.61s
	ALC	0.5642	0.5804	0.5896	0.5938	0.6559	0.4952	0.6543
	NAUC	0.779	0.7852	0.8371	0.9043	0.9321	0.6992	0.9258
	Overfit	Yes	Minor	No	No	No	No	No
EFFICIENTNET-B0	Return Time	78.87s	77.85s	88.02s	85.57s	84.27s	85.77s	84.2s
	ALC	0.5880	0.5831	0.6165	0.6101	0.6542	0.4906	0.6584
	NAUC	**0.8233**	0.7657	**0.90**	**0.9231**	**0.9548**	0.6905	0.9489
	Overfit	Yes	Yes	No	No	No	No	No

TABLE I: Numerical evaluation of different network architectures trained with two fully-connected layers of size 512.

Architecture	Parameters			FLOPS
	Backbone	FullyConn	Total	Total
MOBILENETV2	196K	1'007K	1'203K	7.5M
EFFICIENTNET-MINI	2'303K	1'253K	3'556K	21.8M
EFFICIENTNET-B0	3'595K	1'335K	4'930K	30.1M

TABLE II: Number of parameters and FLOPS used.

	Validation datasets				
	Image		Video		
	Idead	Freddy	Homer	Isaac2	Formula
ALC	0.6808	0.7907	0.2957	0.6131	0.7762
	Test datasets				
	Image		Video		
	Apollon	Loukoum	Fiona	Monica	Kitsune
ALC	0.5776	0.8751	0.4166	0.4103	0.1666

TABLE III: Performance of the final model on hidden data.

sequence length of the video. For short videos ($\leqslant 48$ frames), a temporal kernel size of 3 is chosen and 4 consecutive frames are selected. For longer videos (> 48 frames), the temporal kernel size was 7 and 8 frames are selected with a stride of 4.

Apart from the random key frame selection, all further augmentation techniques slowed down the learning process, leading to worse ALC. The same holds true for more advanced key frame selection techniques during inference (e.g. variation based selection). All images in the videos were resized to 80×80 if they exceeded a size of 90×90.

III. EXPERIMENTS & DISCUSSION

A. MOBILENETV2 *vs.* EFFICIENTNET

EFFICIENTNETS are defined as differently scaled versions of a base model EFFICIENT-B0. They consistently outperform state of the art architectures on ImageNet while using orders of magnitude fewer parameters and FLOPS. This makes them a prime candidate for our search for a fast and powerful architecture. Since the base model is already considerably bigger than MOBILENETV2, we also considered a scaled down version using a scale factor 0.8, which we call EFFICIENTNET-MINI. Table I shows the return time (time from the start of the program until the end of the first test predictions), ALC, as well as the Normalized Area Under the Curve (NAUC). In terms of ALC, MOBILENETV2 clearly is the superior choice, mainly being due to the slow return time of the EFFICIENTNET based models. Overfitting is reported in Table I when the test set accuracy (NAUC) drops continually during training. Interesting to note is that the difference in speed between the EFFICIENTNETS is negligible, even tough there is a significant difference in terms of parameters and FLOPS (Table II). From the final NAUC numbers it is clear that EFFICIENTNET-B0 is the most powerful learner, but falls short in terms of ALC due to speed constraints.

B. Validation and test set results

The performance of the final automated deep learning pipeline on validation and unseen test datasets is presented in Table III. The final model uses adaptive resizing for preprocessing, MOBILENETV2 as backbone, temporal convolutions for video, and three fully connected layers as network head. We achieve the 3^{rd} rank for the validation datasets and an average rank of 8 during the test phase of the challenge.

IV. CONCLUSIONS AND OUTLOOK

Our experiments show that the ALC metric, as imposed by the challenge, is highly focused on speed and therefore discourages the use of more complex, slow models as well as regularization and augmentation. Accordingly, we aimed at improving sample efficiency by using completely unregularised models. Therefore, developing alternative means to address overfitting is a promising future direction. We see the merit of such efficient models primarily in practical industrial applications [3]. Moreover, we extended our computer vision model to video processing using temporal convolutions.

Acknowledgement: This work has been supported by Innosuisse grant 25948.1 PFES "Ada".

REFERENCES

1 Z. Liu, I. Guyon, J. Junior, M. Madadi, S. Escalera, A. Pavao, H. Escalante, W.-W. Tu, Z. Xu, and S. Treguer. Autocv challenge design and baseline results. 2019.

2 M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

3 T. Stadelmann, M. Amirian, I. Arabaci, M. Arnold, G. F. Duivesteijn, I. Elezi, M. Geiger, S. Lörwald, B. B. Meier, K. Rombach, et al. Deep learning in the wild. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 17–38. Springer, 2018.

4 M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.