To be determined

Felipe Augusto Lima Reis, Raquel Almeida, Silvio Jamil F. Guimarães, Zenilton K. G. do Patrocínio Jr
Audio-Visual Information Processing Laboratory (VIPLAB)
Pontifical Catholic University of Minas Gerais (PUC Minas)
Belo Horizonte, Minas Gerais, Brazil
{falreis, raquel.almeida.685026}@sga.pucminas.br, {sjamil, zenilton}@pucminas.br

***Abstract*—**
***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Image segmentation refers to the partition of an image into a set of regions representing meaningful areas. It is considered a challenging semantic task aiming to determine and group uniform regions for analysis. verificar se eh plagioAccording to [1], to create an adequate segmented image it is necessary that the output presents some fundamental characteristics, such as: (i) regions of an image segmentation should be uniform and homogeneous with respect to some features, such as gray level, color, or texture; (ii) region interiors should be simple and without many small holes; (iii) adjacent regions of a segmentation should have significantly different values with respect to the features on which they are uniform; and (iv) boundaries of each segment should be smooth, not ragged, and should be spatially accurate. Segmentation is an active topic of research and in a traditional approach the task is performed using hand-engineered features.

Recently, deep learning architectures drastically changed the computational paradigm for visual tasks. The main advantage of deep learning algorithms is that it does not require an engineered model to operate, meaning that they are capable of learning not only the features to represent the data but also the models to describe it [2]. The success of these approaches relies on a hierarchy of concepts learned through the network, in which more complex concepts are build from simpler ones. In the deep learning approach applied in images, the raw pixel on the input layer is learned as segments and parts until the composition of multiple object concepts later in the network.

Unsurprisingly, many approaches have been proposed to explore these hierarchies, creating maps from the outputs of different layers of a deep learning network. One challenge in this strategy is how to combine these maps, considering that they are presented with different sizes and could represent different concepts. In this work it is presented strategies to combine hierarchical maps to create region proposition for the task of binary image segmentation.

The remainder of this work is organized as bla bla bla....

The authors are grateful to FAPEMIG (PPM 00006-16), CNPq (Universal 421521/2016-3 and PQ 307062/2016-3), CAPES (MAXIMUM STIC-AmSUD 048/14) and PUC Minas for the financial support to this work.

II. Related work

In the earlier years of the deep learning resurgence, an strategy in [3] (extended version in [4]), tackles the task of scene parsing—segmentation task applied for each pixel of the image, aiming to group pixels composing all the identifiable objects in the scene—using hierarchical trees and deep features alongside. Images are used as input for a convolutional network to extract deep features from multiple scales of the images, and in parallel to construct a segmentation tree, to represent in its nodes dissimilarities of neighboring pixels. The tree nodes are used to pool the correspondent deep features to be processed by a classifier. The classifier scores are used to create histogram of object classes for each node of the segmentation tree, and the final parsing proposal is built using the class entropy distribution for selecting the nodes that cover the entire image.

The proposal in [3] with an auxiliary hierarchical structure was one of the first strategies to extend the use of deep features to a complex task. It is important to bear in mind that deep learning approaches were initially described as black-box methods, meaning that not much were known about the reasoning and decisions of the created models. Much exertion have been applied to investigate the networks operation, whether by methodical experimentation [5]–[8] or visualization methods [9]–[11]. Those efforts provided more clarity of the hierarchical aspects of the deep features, which allowed researches to explore these aspects in their endeavors.

In exploring the hierarchies of deep features, three main architectures stand out in recent years, namely: (i) Holistically-nested Edge Detection (HED); (ii) Convolutional Oriented Boundaries (COB); and (iii) Rich Convolutional Features (RCF). Those networks explicit explore the hierarchies by extracting side outputs of traditional convolutional networks to create boundary maps which are also learned in the network.

To the best of our knowledge, the first network exploring this strategy was HED (extended version in [12]), which applied the boundary maps for the boundary detection task, aiming to identify the limits separating uniform regions. The HED network create an side-output layer at each stage of the VGG16 network [13], in which the stages are composed by two Convolution+ReLU layers followed by a Max Pooling layer. In HED, each side-output layer is associated with a classifier in a deeply supervised scheme [14]. The layers create edge maps, which are scaled and fused at the end, to

be evaluate by a cost-sensitive function to balance the bias towards not-boundary pixels. The HED network significantly improved the performance in multiple datasets. The extended version also applied the network for the segmentation task. The authors in [15] use the edge maps created by the HED network alongside with other features such as brightness, colors, gradient and variance to describe images. The goal of their proposal was to create an efficient framework to be used as real-time segmentation system, focused on a fusion strategy to update region features.

In the COB network, the authors also create edge maps from side activations, differing maily from HED by the attribution to candidate contours the orientation information and weights representing the contour strength. The contour orientations are estimated by approximation to known polygon segments and are used to create segmentations hierarchies. The segments weights are computed based on the candidate contour neighboring region to measure the confidence that the candidate is a boundary line. The weights are thresholded to determine the granularity of the segment when creating the segmentation hierarchy. The network perform well in multiple tasks such object proposal, object detection, semantic contour and segmentation.

Finally, the RCF network applied in the boundary detection task, which differ from HED by three main modifications. The first regards the input layer, in which it is used pyramids to create multiple scales of the images. The scaled images are later interpolated in the output layer, similar to [3]. The second modification regards the number of side output maps. RCF creates a side output at each Convolutial+ReLU layer of the VGG16 network, which is believed to create more detailed representations and improve the network accuracy. The last modification is in the loss function and the ground-truth of the datasets. In the ground-truth images, pixels are weighted based on a vote among multiple human-annotated values. Any pixel that not achieve a confidence vote value is disregarded by the loss function in the network. The goal is to reduce inconsistencies in the fallible human annotations and mitigate the network confusion in controversial pixels.

In [16] the authors pursued a similar direction of the afore mentioned networks. The proposal consists of joint strategies for the recognition task in large scale, specifically: NAO ENTENDI.

III. Hierarchical maps in convolutional neural networks

This work present strategies to merge hierarchical maps created from outputs of different layers of a convolutional network.verificar se eh plagio. In a convolutional network each layer is a three-dimensional array of size $h \times w \times d$, where h and w are spatial dimensions and d is the feature, channel or stride dimension. The first layer is the input image, with pixel size $h \times w$ and d color channels. Locations in higher layers correspond to the locations in the image they are path-connected to, which are called their receptive fields. Convolutional networks are built on translation invariance and their basic components (convolution, pooling, and activation functions) operate on local input regions and depend only on relative spatial coordinates.

The convolutional network model used in this work is the VGG network [13],proposed in 2014 as one of the first attempts to create deeper models for the task of object recognition. The architecture is a composition of multiple stacked convolutional layers, in which the receptive fields and stride have a fixed $3 \times 3 \times 1$ size. Following each two or three layers of convolution is placed a max-pooling layer. Also, all hidden layers are supplied with a ReLU non-linear rectification.

As demonstrated in bla bla layers bla bla hierarchies bla bla for binary segmentation bla bla bla

Formally, let $S = \{(X_n, Y_n), n = 1, ..., N\}$ be the training input set for the network, in which X_n is a set of N images with three color channels and Y_n the set of N labels associated with each image with values belonging to $\{0, 1\}$. Consider also $\mathbf{W}$ the layer set of parameters in which $\mathbf{w} = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$ is the associated weights for each one of the M side output maps. The objective function for training the weights for the ℓ_{side} image map could be defined as:

$$\mathcal{L}(\mathbf{W}, \mathbf{w}) = \sum_{m=1}^{M} \alpha_m \ell_{side}^{(m)}(\mathbf{W}, \mathbf{w}_m) \quad (1)$$

IV. Experiments

bla bla bla

A. The Kitti dataset

KITTI Vision Benchmarking Suite is a project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago to provide a real-world computer vision benchmark for autonomous driving platform Annieway. KITTI contains benchmarks and datasets for the following area of interests: stereo, optical flow, visual odometry, 3D object detection and 3D tracking.

One of the benchmarks in the KITTI suite is the Road/Lane Dataset Evaluation [17]. The road and lane estimation benchmark consists of 289 training and 290 test images, in four different categories of road scenes [17]:

- uu - urban unmarked (98 training images and 100 test images) [17];
- um - urban marked (95 training images and 96 test images) [17];
- umm - urban multiple marked lanes (96 training images and 94 test images) [17];
- urban - combination of the three above [17].

Ground truth has been generated by manual annotation of the images and is available for two different road terrain types: road - the road area (the composition of all lanes), and lane (the ego-lane, the lane the vehicle is currently driving on) [17] . Ground truth is provided for training images only [17].

As the dataset does not provide test ground truth, the results must be evaluated using a benchmarking tool provided with the dataset. This tool performs road and lane estimation in the birds-eye-view space [17] . The metrics used are

Maximum F1-measure, Average precision as used in PASCAL VOC challenges, Precision, Recall, False Positive Rate, False Negative Rate, F1 score and Hit Rate [17].

B. Experimental setup

C. Results

D. Qualitative analysis

REFERENCES

[1] D. Domnguez and R. R. Morales, *Image Segmentation: Advances*, 2016, vol. 1, no. 1.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1, 2016.

[3] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Scene Parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers," in *29th International Conference on Machine Learning (ICML 2012)*, A. McCallum, Ed., Edinburgh, United Kingdom, Jun 2012, pp. 1–8.

[4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, Aug 2013.

[5] R. Ilin, T. Watson, and R. Kozma, "Abstraction hierarchy in deep learning neural networks," in *International Joint Conference on Neural Networks, 30.* Anchorage, Alaska: IEEE Computer Society, 2017, pp. 768–774.

[6] C.-C. J. Kuo, "Understanding convolutional neural networks with a mathematical model," *Journal of Visual Communication and Image Representation*, vol. 41, pp. 406–413, 2016.

[7] D. Eigen, J. Rolfe, R. Fergus, and Y. Lecun, "Understanding deep architectures using a recursive convolutional network," in *International Conference on Learning Representations*. Banff, Canada: Computational and Biological Learning Society, 2014.

[8] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations*. Toulon, France: Computational and Biological Learning Society, 2017.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, 2013.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, vol. 8689. ZURICH, Switzerland: Springer International Publishing, 2014, pp. 818–833.

[11] B. Alsallakh, A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Do convolutional neural networks learn class hierarchy?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 152–165, 2018.

[12] S. Xie and Z. Tu, "Holistically-nested edge detection," *International Journal of Computer Vision*, vol. 125, no. 1, pp. 3–18, Dec 2017.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[14] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-Supervised Nets," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Lebanon and S. V. N. Vishwanathan, Eds., vol. 38. San Diego, California, USA: PMLR, 09–12 May 2015, pp. 562–570.

[15] M.-M. Cheng, Y. Liu, Q. Hou, J. Bian, P. Torr, S.-M. Hu, and Z. Tu, "HFS: Hierarchical Feature Selection forEfficient Image Segmentation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 867–882.

[16] J. Fan, T. Zhao, Z. Kuang, Y. Zheng, J. Zhang, J. Yu, and J. Peng, "HD-MTL: Hierarchical Deep Multi-Task Learning for Large-Scale Visual Recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1923–1938, Apr 2017.

[17] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

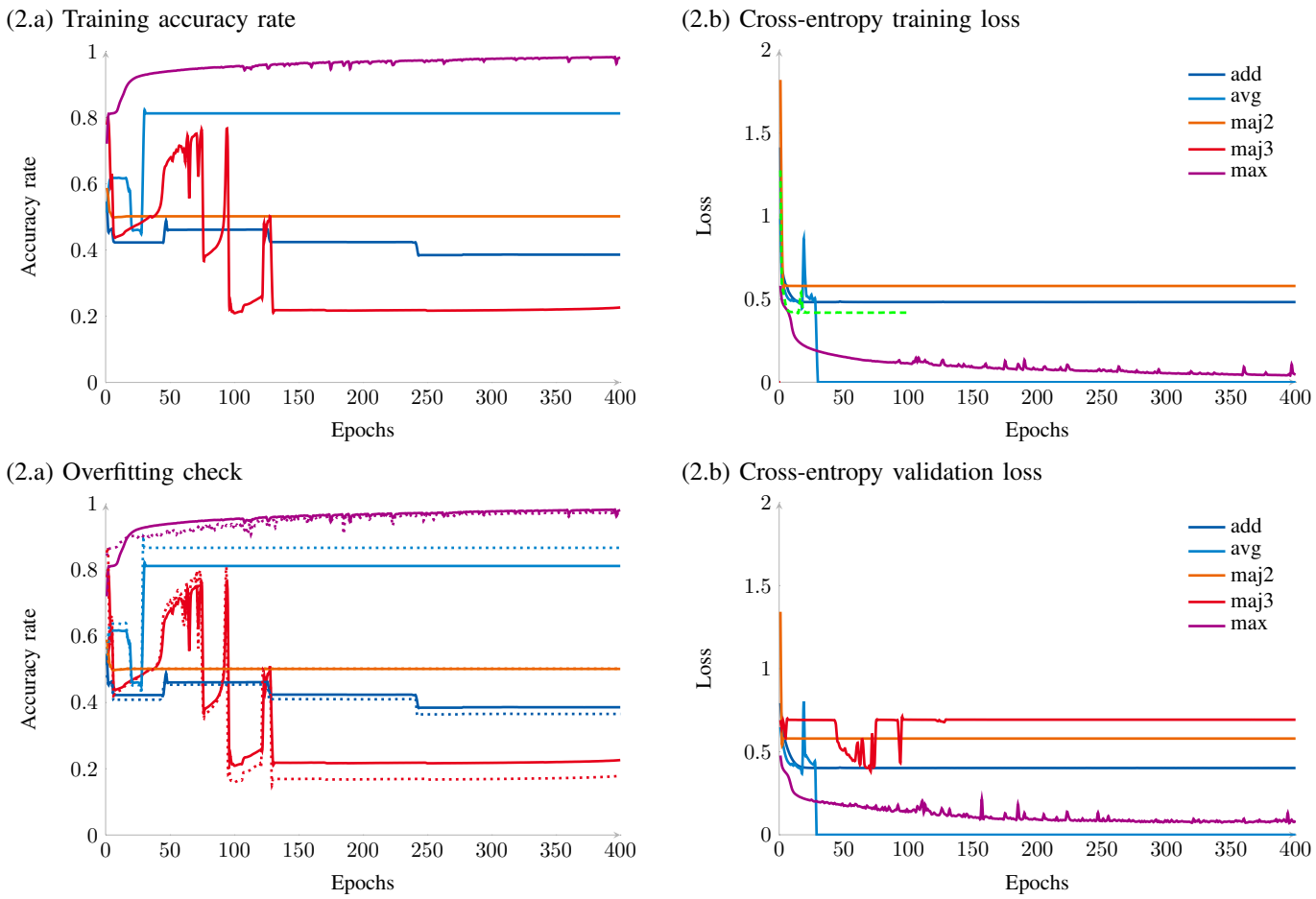

Fig. 1. Aug - Learning curves for the compared approaches. Left panel displays the accuracy obtained on the training and validations sets. Right panel displays the cross-entropy objective function

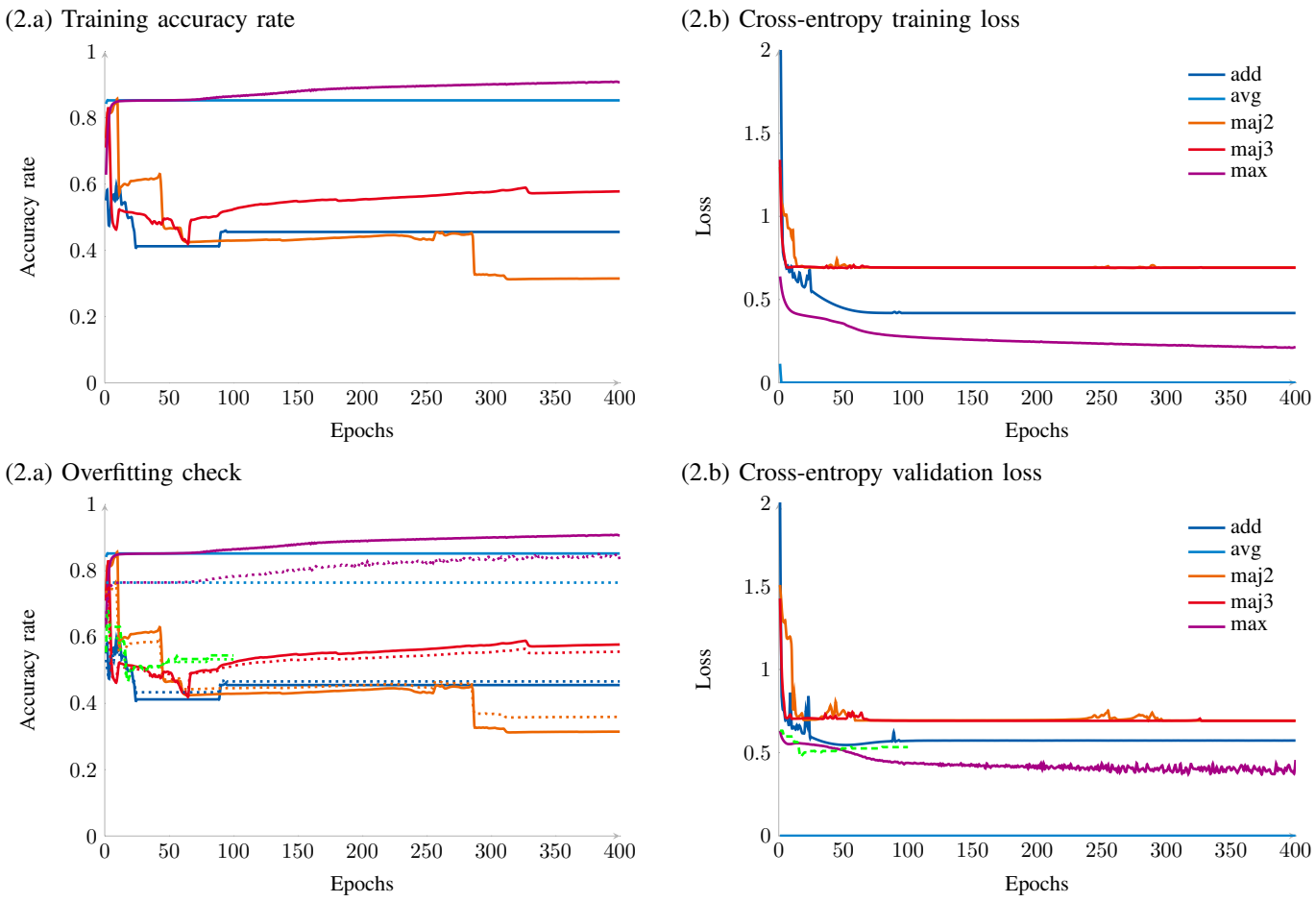

Fig. 2. No Aug - Learning curves for the compared approaches. Left panel displays the accuracy obtained on the training and validations sets. Right panel displays the cross-entropy objective function