Convoluted Mixture of Deep Experts for Robust Semantic Segmentation

Abhinav Valada*†, Ankit Dhall*‡ and Wolfram Burgard†

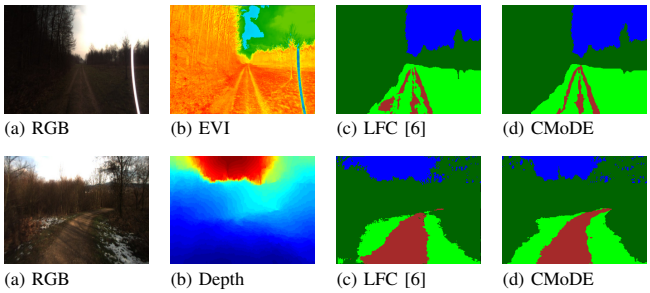
***Abstract*— Robust scene understanding of outdoor environments using passive optical sensors is a critical problem characterized by changing conditions throughout the day and across seasons. The perception models on a robot should be able learn features impervious to these factors in order to be operable in the real-world. In this paper, we propose a convoluted mixture of deep experts (CMoDE) model that enables a multi-stream deep neural network architecture to learn features from complementary modalities and spectra that are resilient to commonly observed environmental disturbances. Our model first adaptively weighs features from each of the individual experts and then further learns fused representations that are robust to these disturbances namely shadows, snow, rain, glare and motion blur. We comprehensively evaluate the CMoDE model against several other existing fusion approaches and show that our proposed model exceeds the state-of-the-art.**

I. INTRODUCTION

As robots are progressively being deployed for real-world outdoor tasks, scene understanding plays a pivotal role for successful operation. Perception in outdoor environments is inherently more challenging than indoors, considering the frequent appearance changes that take place due to the varying environmental conditions. Some of these disturbances such as shadows cause a minor change in appearance that can be filtered or transformed to an invariant color space [4], whereas others such as snow, rain, low-lighting and motion blur have adverse effects that are hard to negate. Features from complementary modalities and spectra that are not influenced by these disturbances can be intelligently fused to obtain a robust feature set.

Recent advances in Deep Convolutional Neural Networks (DCNNs) and new multimodal datasets have significantly improved the state of the art of various robotic perception problems. Eitel *et al.* [1] use a late-fusion approach for object detection, where two individual networks are first trained with RGB and depth data respectively and their features are concatenated to yield a combined prediction. Valada *et al.,* [7] use a similar approach for semantic segmentation, but in addition of combining features from multiple networks, their model further learns fused filters using a stack of convolution and pooling layers. Hinton *et al.,* [2] introduced the classical Mixture of Experts (MoE) model, where experts map the input to a set of outputs and a gating network produces a probability distribution over the experts. Whereas, Mees *et al.,* [5] extended the MoE to detect people in varying illumination. Their gating network uses inner-product layers to yield two scalars, that are then weighted over each expert to obtain a combined prediction.

*These authors contributed equally. †University of Freiburg, Germany. ‡VIT University, India. This work has partly been supported by the European Commission under FP7-267686-LIFENAV and FP7-610603-EUROPA2.



(a) RGB (b) EVI (c) LFC [6] (d) CMoDE

(a) RGB (b) Depth (c) LFC [6] (d) CMoDE

Fig. 1. Qualitative comparison of the segmentations obtained using the LFC [6] and the CMoDE approach. The first row shows the segmentation in highly shadowed areas with glare ($g_{rgb}(X_1) = 0.49,\ g_{evi}(X_1) = 0.51$), while the second row shows the performance in the presence of snow ($g_{rgb}(X_2) = 0.47,\ g_{depth}(X_2) = 0.53$).

In contrast to these approaches, we propose a Convoluted Mixture of Deep Experts (CMoDE) model that builds upon [5], to fuse multiple modalities or spectra for semantic segmentation. The model has two components: experts that map particular modalities to the segmentation output and the adaptive gating network that learns "how much" and "when" to rely on each expert. We train the network to learn the convex-combination of the experts by back-propagating into the weights, similar to any other synapse weight or convolutional kernel. We show that our approach exceeds the state-of-the-art fusion techniques, in addition to demonstrating robust segmentation in adverse environments.

II. CONVOLUTED MIXTURE OF DEEP EXPERTS

We represent the training set of a CMoDE, with E experts and C segmentation classes, as $S = \{(X_n, y_n),\ n = 1, 2, \ldots, N\}$. Each training example, $X_n = (x_1, x_2, \ldots, x_E)$ is a vector of raw images from different modalities or spectra, where image x_i is shown only to the i-th expert. y_n is a $W \times H$ segmentation mask, where, $y_n(r,c) \in \{0, 1, \ldots, C\}$, for $r,c \in \{1, 2, \ldots, W\} \times \{1, 2, \ldots, H\}$, maps the membership of pixel $X_i(r,c) \in X_n$ for each modality, in one of the C classes. Since experts trains on images from different modalities or spectra, each one specializes in a particular sub-space of X_n. The i-th expert, produces it's own segmentation mask, denoted by $h_i(x_i)$. The final segmentation mask is a convex combination of the outputs of E experts; weighted by $g(X_n)$, which is the output of the adaptive gating network. Our network further learns fused representations over these outputs using a convolution layer. The final output is a per-pixel segmentation mask $\hat{y}_n$, corresponding to the input X_n and is written as,

$$\hat{y}_n = f(X_n) = softmax\left(\mathcal{W} * \left[\sum_{i=1}^{E} g_i(X_n)\, h_i(x_i)\right]\right) \qquad (1)$$

where, $g_i(X_n)$ corresponds to the scalar weight for the i-th expert and $\mathcal{W}$ is a stack of convolution kernels learned on

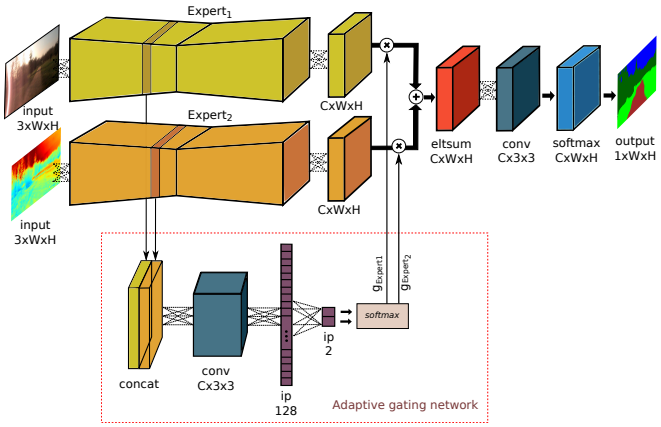


Fig. 2. Convoluted Mixture of Deep Experts architecture configuration. Any DCNN based segmentation network can be plugged in for each of the experts depicted. We use the DCNN described in [7] for our experiments.

the fused representation and $*$ is the convolution operation. Replacing the input to the gating network, consisting of raw pixels; with a representation of X_n from within the expert network defined by $\rho(X_n) = (r(x_1), r(x_2), \ldots, r(x_E))$, yields improved results with lesser computation time. $r(x_i)$ is a representation of x_i taken from the i-th expert, for instance, the output of *conv4*. Defining r from the contracting part of the expert network leverages the fact that W and H decrease, while channel depth increases towards the end of the contracting part, forcing the network to increase the "what" and reduce the "where" [6]. This "what" is of primary importance to the adaptive gating network. For instance, if an RGB image is washed out due to poor lighting conditions, the network needs to only know "what" and not "where" the image is washed out; making it rely less on the RGB expert and give it a lower score, $g_{RGB}(X_n)$, while relying more on other experts. Re-writing the equation with $\rho(\cdot)$,

$$\hat{y}_n = f(X_n) = softmax\left(\mathcal{W} * \left[\sum_{i=1}^{E} g_i(\rho(X_n)) \cdot h_i(x_i)\right]\right) \quad (2)$$

Figure 2 depicts the generic architecture of our network: the DCNN experts and the adaptive gating network. We use two experts for simplicity, but the architecture is generalizable for arbitrary number of experts. Each expert network is trained separately and uses a subspace of X_n to train. For this work, we use the base architecture described by Valada *et al.* [7] for each expert. The adaptive gating network takes $\rho(X_n)$ as input and produces probability values to weight each expert. The 3D volumes each of size $C \times W \times H$ from each expert are weighted according to $g(\rho(X_n))$. Convolutions followed by a *softmax* layer converts these to per-pixel class membership probability. It should be noted that while training the adaptive gating network the weights of the experts are kept constant.

III. EXPERIMENTAL RESULTS

We use the publicly available Freiburg Multispectral Forest dataset [7] for our experiments and the Caffe [3] deep learning framework for the implementations. We consider three different modalities and spectra, namely, RGB, depth and Enhanced Vegetation Index (EVI). EVI, which is computed using RGB and Near-infrared data, is useful for highlighting high biomass regions and vegetation monitoring. Recent work has demonstrated the utility in using EVI with RGB for segmentation in forested areas [7].

TABLE I

COMPARISON OF DEEP FUSION APPROACHES.

Input	Approach	IoU	PA	FPR	FNR
RGB	Unimodal	84.90	94.47	7.80	7.40
DEPTH	Unimodal	76.10	88.93	12.76	11.14
EVI	Unimodal	83.25	93.28	8.70	8.10
RGB-D	Average	80.75	91.45	10.83	8.42
	Late-fused Conv [7]	84.04	93.19	9.40	6.55
	CMoDE	86.79	93.92	7.17	6.04
RGB-E	Average	84.88	93.75	8.29	6.83
	Late-fused Conv [7]	86.90	94.44	7.00	5.76
	CMoDE	86.97	94.49	7.12	5.91

Tab. I shows the performance comparison of our CMoDE model with different fusion approaches. The metrics shown (in %) correspond to Mean Intersection over Union (IoU), Mean Pixel Accuracy (PA), False Positive Rate (FPR), False Negative Rate (FNR). For a baseline, we also show the performance obtained while averaging predictions from each network trained on a specific modality. It can be seen that averaging demonstrates a poor performance compared to other techniques. Our CMoDE model trained on RGB and depth images, yields a 6.04% improvement over the averaging approach and 2.74% improvement over the Late-fused Convolution (LFC) approach. While using RGB and EVI as input, the CMoDE model yields a comparatively better performance and demonstrates better qualitative results as shown in Fig. 1. The first row shows the segmentation in low-lighting and glare, while the second row shows results in the presence of snow. It can be seen that the CMoDE model accurately segments the scene in the presence of these disturbances. In addition, a live demo can be accessed at `http://deepscene.cs.uni-freiburg.de/`.

IV. CONCLUSION

We introduced a new deep adaptive fusion architecture for end-to-end segmentation of multimodal and multispectral images. Our CMoDE model achieves state-of-the-art performance compared to unimodal segmentation and existing fusion approaches. More importantly, the model demonstrates considerable robustness to commonly observed environmental disturbances, critical for real-world robotic perception.

REFERENCES

[1] A. Eitel et al., "Multimodal Deep Learning for Robust RGB-D Object Recognition", Int. Conf. on Intelligent Robots and Systems, 2015.

[2] R. A. Jacobs, M. I. Jordan, S. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts", Neural Computation, 3:112, 1991

[3] Y. Jia et al., "Caffe: Convolutional Architecture for Fast Feature Embedding", arXiv preprint arXiv:1408.5093, 2014.

[4] W. Maddern et al., "Illumination Invariant Imaging: Applications in Robust Vision-based Localisation, Mapping and Classification for Autonomous Vehicles", ICRA workshop, 2014.

[5] O. Mees et al., "Choosing Smartly: Adaptive Multimodal Fusion for Object Detection in Changing Environments", IROS, 2016.

[6] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", MICCAI, 2015.

[7] A. Valada et al., "Deep Multispectral Semantic Scene Understanding of Forested Environments using Multimodal Fusion", ISER, 2016.