UDC: 004.932 https://doi.org/10.15407/jai2023.01.074

INTELLIGENT RECOGNITION AND INTEGRATION OF GRAPHICAL ELEMENTS INTO VIRTUAL SURROUNDING WITHIN AUGMENTED REALITY USING HYBRID CONVOLUTIONAL NEURAL NETWORKS

V. Sineglazov[1], **I. Boryndo**[2]
[1,2]National Aviation University, Ukraine
1, Lubomyr Husar ave., Kyiv, 03680
svm@nau.edu.ua
ibo.mistle@gmail.com
[1]https://orcid.org/0000-0002-3297-9060
[2]https://orcid.org/0000-0001-5375-6272

Abstract. In this paper analysis of modern augmented reality algorithms based on mobile devices was done. As a result, algorithmic shortcomings were identified and the usage of convolutional neural networks was proposed. Within the research the qualitative analysis of modern architectures of convolutional neural networks was carried out and their separate shortcomings at use in systems on the basis of processor architecture ARM was shown. As a result of this research it was found that to achieve the target accuracy and speed of the system it is important to use a hybrid convolutional neural network, which significantly improves the quality criteria of the system. The optimal structure and parameters for initialization and training of a hybrid convolutional neural network system used for augmented reality are obtained. The optimal training sample was formed and the use of pre-trained HCNN on another device of ARM architecture was described.

Keywords. Convolutional neural network, augmented reality, machine learning, virtual reality, image classification, ARM.

Introdaction

In today's world, augmented reality technologies are becoming more common every year, and the demand for their more flexible technological implementation is only growing. Currently, augmented reality functions have many types and can be both separate specialized applications and functional additions to existing ones. For their identity, augmented reality systems strive for high accuracy and efficiency in processing environmental data, the user's position in space and as a result of clearly fitting target objects into space with minimal error. This requires a clear algorithm for evaluating graphical data. Algorithms for soleistic and polygonal processing of environmental photographs are commonly used, but they are not flexible in configuration and work poorly in environments with different lighting conditions. To solve these problems, in this paper we're proposing to consider a system of several neural networks.

Currently, there are a large number of image processing tasks and augmented reality is one of them. To solve them, the use of convolutional neural networks is the key, but a number of criteria should be considered, such as the usage platform, target accuracy and system performance. Most modern types of convolutional neural networks are severely limited due to their speed and complexity of learning, and at the same time they require a complete quality training sample. Over the years, convolutional neural network architectural complicity growing in order to solve such problems and improve result quality and accuracy which leads to new problems when further structural enrichment of convolutional neural networks encounters hardware limitations. In such conditions, the use of hybrid convolutional neural networks becomes crucial. To increase overall performance and accuracy, multiple convolutional neural networks can be combined into a single system to form a hybrid convolutional neural network. In this paper we will describe and present the results of research on the features of use and topology of hybrid convolutional neural networks, their optimization and training.

Applications of convolutional neural networks within augmented reality

Augmented reality by itself is the complex of interrelated different algorithms of image recognition, processing and enrichment to add believable graphical elements or features to surrounding reality. Since, the main process of augmented reality implementation is based on surrounding recognition that is the raw graphical data by itself, it means that recognition process can be done using convolutional neural networks. Augmented reality systems could be divided into different types that in result brings up different targeted tasks that should be solved using neural networks. The main types of augmented reality are:

– surrounding augmentation-based reality (it based on recognition and enhancing the graphical data of subjects surrounding in real time via pocket device/inserted camera, etc.);

– procedurally generated virtual reality (unlike the enhancing of real-life data it based on generation of whole new reality based on the either preloaded graphical data or real time surrounding recordings using VR devices).

Based on type of augmented reality and its project specific, it's possible to extract the recognition task that could be solved through convolutional neural network. As an example, let's consider augmented reality that transfers surroundings to whole VR space using VR devices.

It means that 4 targeted cameras records nearest space in 240 degrees cone in front of a user and creates polygonal map that can be mapped and textured via VR space application.

Therefore, based on the control type (controller, gestural, verbal, etc.) the following tasks could be highlighted:

– processing of camera outputs to get following environmental parameters: position of the user relatively to ground, sight vector, lightning sources, movement type (still, crouching, jumping, tilting, walking), etc;processing of hands gestures using frontal camera data: to analyze hands or fingers position, forms, and movement style to recognize predefined gestures. It could be applied as a way of controlling augmented reality features;

– extracting the specific places in the surrounding that suits to further insertions there of some object that could be both sprite or 3D model;

– for object insertion task, based on the information about insertion place (lighting, special characteristics, distance), to apply to target object logical adjustments such as resize, recolor, shadowing, perspective shifting to make it suitable and realistic considering the environment;

– recognize other people in the line of sight, their poses and movements.

Fig. 1. Example of hands gesture recognition for augmented reality application as the way to interact or control virtual functionality

Topology analysis of modern convolutional neural networks

Before we'll dive in the details of practical application of hybrid CNN systems for augmented reality implementation it's necessary to take a look onto CNNs itself. Nowadays convolutional neural network (CNN) is the basic tool for graphical data processing and feature extracting. It's a notable deep architecture of deep learning. Due to its specific structure these neural networks can automatically extract the features or representation characteristics by processing the number of prepared graphical input data. Basically, one part of CNN is extracting features and another is processing them and classifying due to initial task requirements.

In the same way hybrid convolutional neural network essentially is the combination of two or more different convolutional neural networks that configured and structured to work in pair (parallel or serial) and solve specific task or range of tasks.

The idea of HCNN architecture is to implement the single-responsibility principle that makes each component of our systems (e.g. CNNs, classification/recognition algorithms, input/output data processors) to perform only one specific task. Therefore, we can decompose complex specific tasks into required dataflow steps that should be applied. Each step will be processed with its specific element.

Having these simplified tasks its makes easier to train responsible neural networks and increases its accuracy and performance. As the example following networks can be combined: so called densely connected convolutional neural network in pair with squeeze and excitation convolutional neural network based onto ResNeXt. It has a good potential due to global information holding at SE-CNN structure and DenseNet performance results. While combining networks the number of parameters should be considered:

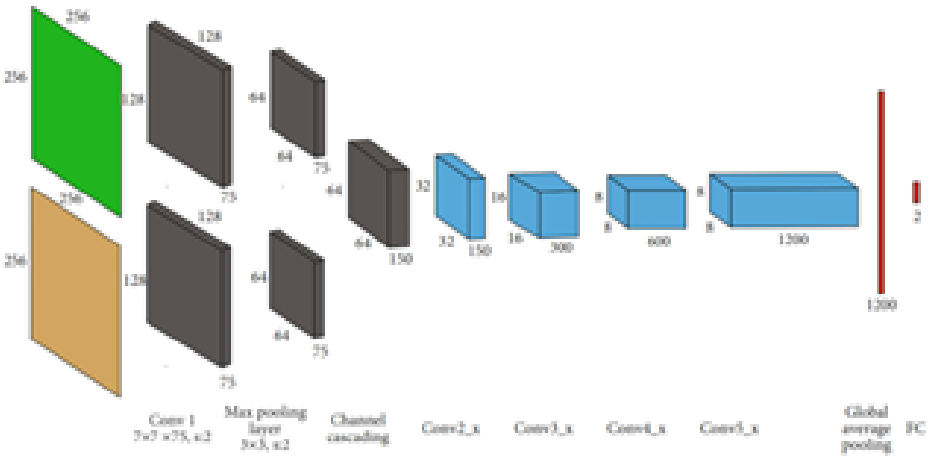

Fig. 2. Hybrid convolutional neural network simplified structural scheme

– input data parameters such as initial scale, resolution, number of channels and recognition task type;

– the output of first CNN should be acceptable by the second one, initial target information should be saved;

– the structures of both CNNs should beflexible and able to include supportive layers such as normalization layers, residual blocks, dropout layers, 1x1 convolution layers, etc.

In such approach any type of CNN can be paired. It's based mostly on the specifics of tasks and input data parameters (e.g. image resolution, scale, color channels, number of training samples, etc.).

Choosing the optimal mobile-type hybrid convolutional neural network for augmented reality implementation

As the augmented reality is the technology that mostly uses on mobile or VR devices, it's neccessary to choose that type of convolutional neural network type and architecture that could be optimized and easily executed on the devices with limited resources. As an example we're chosing the mobile devices on the basis of ARM architecture with 6-cores of 2.2(1.8)GHz of nominal frequencies, 4GB of LPDDR4X RAM and Adreno 540 GPU. The potential CNN structures shoud be not over complicated with number of layers and contain strong feedback interconnections within different structural levels. Within the list of modern convolutional neural networks there exists a few predefined structures that could be applied for our purposes:

– EfficientNet B0 & EfficientNet B3;
– MobileNet CNN & MobileNet V2 CNN;
– InceptionResNet V2 CNN;
– DenseNet201 & DenseNet169 & DenseNet121;
– Channel-boosted 2OR CNN;
– ResNet101;

So lets pre-train following list of neural networks using CIFAR-100 learning sample and measure the different performance and accuracy criterias. The number of parameters like overall memory usage or pre-trained model file size are cruicial while considering about hardware limitations of usage platform. All the metrics are obtained using python 3.9.0 with Keras framework API.

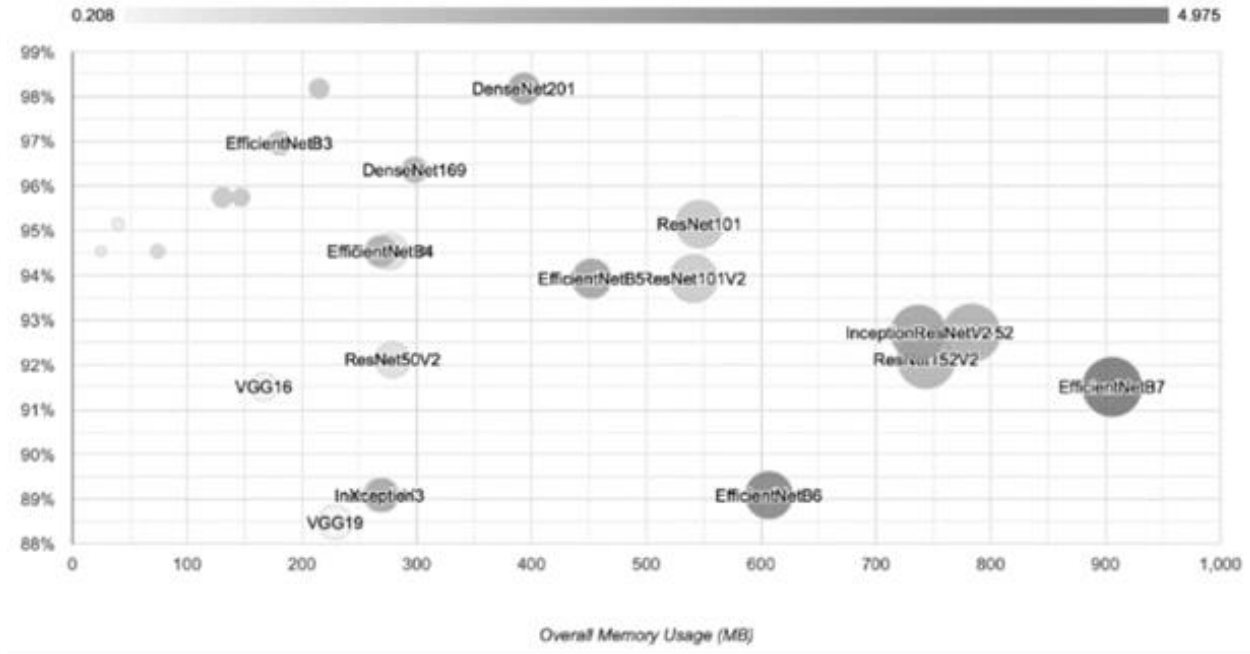

Fig. 3. Memory usage grapgh of modern convolutional neural network architectures

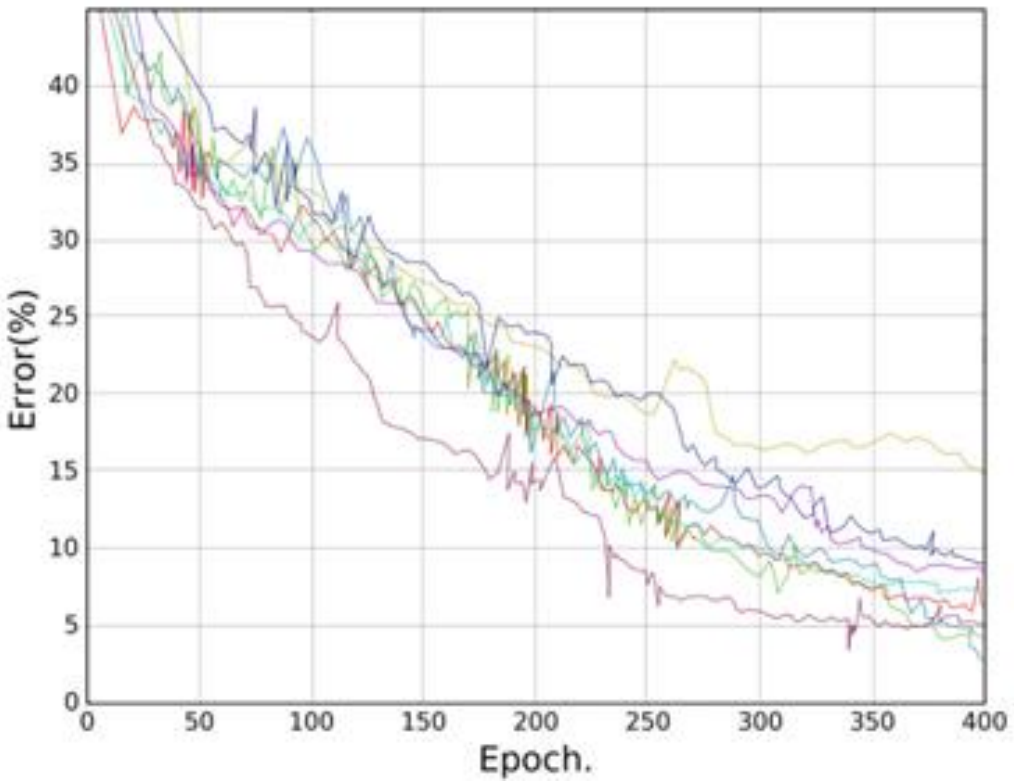

Fig. 4. Training curves for each of convolutional neural networks architecture that were used at performance testing

Table 1. Modern CNN architectures performance analysis based on execution criteria (CIFAR-100)

CNN	Top-1 Accuracy (%)		Inference Time (ms)		Overall Memory Usage (MB)		Model File Size (KB)	
	TF	TFL	TF	TFL	TF	TFL	TF	TFL
DenseNet121	98.18	98.18	1746	168	214.8	46.25	28249	27289
DenseNet169	96.36	96.36	2251	208	298	67.87	50521	48949
DenseNet201	98.18	98.18	2727	252	393.3	90.37	72907	70889
EfficientNetB0	94.54	94.54	1127	145	74.3	26.68	16348	15721
EfficientNetB3	96.96	96.96	1589	273	180.2	62.62	42894	41830
MobileNet	95.15	95.15	0415	182	40	29.76	12900	12547
MobileNetV2	94.54	94.54	0648	29	25	17.42	9271	8713
InceptionResNetV2	92.72	92.72	2762	329	737	248.18	213611	212262
ResNet101V2	93.93	93.93	1457	334	541.1	191.81	167343	166253

Based on the obtained result in table 1, it was possible to identify MobileNetV2 as the model with the lowest complexity, an expected result since it was developed to be efficient specifically for mobile and embedded vision applications that have limited memory and computational power. Also, MobileNetV2 was the model with the lowest overall memory usage for running both TF and TFL models. It is important to notice that this architecture is comparable only to two EfficientNet variations, being the remaining architectures 5 to 10 times memory-eager.

The Top-1 accuracy rate remained the same through the conversion between TF and TFL for all architectures, which was expected since a difference would mean implementation disparities, preventing the remaining metrics comparisons. Two models had the highest accuracy: DenseNet121 and DenseNet201 with 98.18%. We would like to notice that the accuracy is application-dependent, so a lower accuracy does not mean that the architecture should be discarded.

Inference time was the only measured parameter that had a significant difference between the TF model that was run on a

desktop and the TFL model that was run on a mobile device. The difference was due to architectural differences between the environments creating a discrepancy between the latency of different operations. The TF model with the lowest inference time was the MobileNet and the TFL model with the lowest inference time was the MobileNetV2.

Finally, after choosing the optimal convolutional neural network architecture it's necessary to form enriched sample of training images. This sample should contain mapped images of e.g. different hand gestures with different but clear light sources. All the images should be the same size and match the resolution of target device cameras (for oculus quest 2 it's 720p resolution). The sample should contain not less then 800 different pictures.

Conclusions

In this paper we've done the review of modern implementation ways of augmented reality in pair with virtual reality devices, common smartphones, etc. In the result of augmented system analysis, the main sub-tasks were excluded. Based on the criteria that is related to the type of augmented reality, it's control mode and usage approach, the classification list of recognition tasks and their implementation details was listed. To solve them we're recommended to use convolutional neural network systems as it's the best tool for image recognition and processing at this moment. To apply this neural network for image processing, the basic review of convolutional neural network essentials was considered, including resulting structural idea, the number or core layers and their configuration parameters. Based on this, there was formed the list of suitable neural network architectures that are potentially usable and applicable for mobile devices on the base of ARM architecture. Pretrained models of such networks were ran through performance testing phase and the results were organized in Table 1. In the result the optimal type of CNN architecture for augmented reality processing using mobile or VR devices were proposed (DenseNet201 or MobileNet V2).

References

1. Iveta Mrazova, Marek Kukacka. "Hybrid convolutional neural networks" IEEE International Conference on Industrial Informatics 10.1109/INDIN.2008.4618146.
2. Chaitanya Nagpal and Shiv Ram Dubey. "A Performance Evaluation of Convolutional Neural Networks for Face Anti Spoofing".
3. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 675–678.
4. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.
5. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440..
6. Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu. "Squeeze-and-Excitation Networks".
7. Iveta Mrazova, Marek Kukacka. "Hybrid convolutional neural networks" IEEE International Conference on Industrial Informatics 10.1109/INDIN.2008.4618146.
8. Chaitanya Nagpal and Shiv Ram Dubey. "A Performance Evaluation of Convolutional Neural Networks for Face Anti Spoofing".
9. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 675–678.
10. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.
11. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440..
12. Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu. "Squeeze-and-Excitation Networks".
13. C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in ICCV, 2015.
14. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in ICCV, 2015.

The article has been sent to the editors 14.11.22.
After processing 10.12.22.