Journal of Physics: Conference Series

Ar-powered educational application based upon convolutional neural network

To cite this article: O M Gushchina *et al* 2019 *J. Phys.: Conf. Ser.* **1278** 012036

Ar-powered educational application based upon convolutional neural network

O M Gushchina[1], A V Ochepovsky[1], N N Rogova[1] and A A Pupykina[2]

[1]Institute of Mathematics, Physics and Information Technology, *Togliatti State University*, 14 Belorusskaya Street, Togliatti, 445020, Russian Federation
[2]Dipartimento di Elettronica e Informazione, *Politecnico di Milano*, Via Ponzio 34/5, Milano, 20133, Italy

[1]E-mail: g_o_m@tltsu.ru

Abstract. This research aims to justify the importance of using convolutional neural networks (CNN) in AR-powered educational applications to provide the learning process with impressive visualization of studied objects. Demonstration of the developed application proves neural network to be its basic part that allows for high-quality accuracy of object localization and image recognition in AR systems. As these systems provide a more informative and impressive way of learning, the application can be used for data mining and big data analysis disciplines to demonstrate real-time operation of localization, clustering and segmentation algorithms.

1. Introduction

Recent advances in computer processing of data, new 3D technologies and efficient imaging technologies introduced a new aspect of real-time systems performing intense computing operations – augmented reality (AR) systems designed to process immense amount of input data and visualize it into learning material when put into educational context [1]. Two most common scenarios that involve AR systems are tracking the dataflow and displaying content in case of efficient recognition of the object; external services are then to be engaged to provide supplementary information which can make the learning material more informative and impressive.

Our goal is to demonstrate the techniques of developing AR-powered educational application based upon marker and markerless technologies that involve using convolutional neural networks.

2. Rationale and Objective

The objective of the research is to justify the importance of convolutional neural networks used for development of AR-powered educational applications (figure 1). Major programming components required are tracking and visualization.

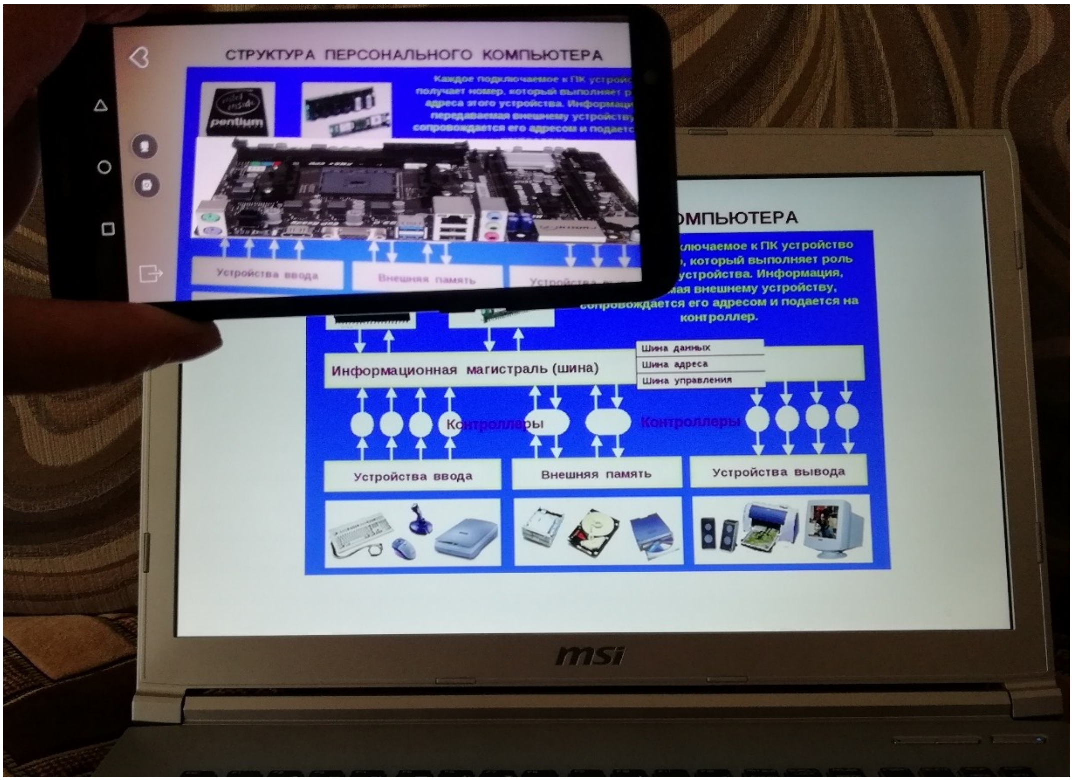


Figure 1. An operating AR application.

AR is based upon a complex of technologies that provide a certain level of its interaction with the real world by aligning objects we perceive conventionally with the virtual ones belonging to virtual reality [2]. Apart from realistic visualization of virtual objects, AR implementation requires their realistic integration into the real world environment. To achieve this, it is necessary to recognize the real object, binding some virtual content to it subsequently; do the rendering, i.e. displaying the virtual content on top of real objects in the image depending on results of recognition; provide their interaction [3]. The latter requires development of an interactive model to allow for additional interactivity. To solve this task, the AR system is provided with computer vision algorithms that find objects and/or special markers in the video stream by highlighting key features of the visual scene (corners and boundaries), searching the objects real-time, performing 3D reconstruction on basis of several photos etc. The way the algorithm performs the search of the object in the video is basically searching it in the static image. As a video file is a sequence of frames (images), this search is repeated throughout a certain number of frames. As the marker is identified and located in the video stream, a projection matrix can be built and virtual models positioned. These can be used to superimpose a virtual object on the video stream in compliance with the angle and perspective. The main challenges here are identifying the marker, locating it in the frame and projecting a compatible virtual model.

AR system tracks codes or markers to recognize real objects and organize their real-time interaction with virtual ones in full compliance with the surroundings. Consider, however, that there are no QR, barcodes or other kinds of markers – how can the algorithm of building AR system be implemented?

A prospective solution of the problem is using neural network [4] able to sequentially locate and recognize objects in the image and then process the visual data. Object localization involves distinguishing it from the surroundings in the image and defining its coordinates, i.e. scanning the image of a given scene for a 2D function domain where this function complies with the model's equivalent one. To identify image features the abstractions of image information are computed and key features such as isolated points, curves and various connected domains are identified.

Besides, developing an advanced AR-powered educational application is not merely creating a neural network: the network must learn and analyze large libraries of objects and output the information according to the specific task set by an educational course.

3. Methods

AR is primarily concerned with identification of real objects and scenarios to supplement them with virtual objects in a robust and realistic way. According to semantic approach, the optimal solution is a technology that uses convolutional neural network (CNN) as a basis to identify and segment objects and scenarios in one or multiple video frames. This process involves optical recognition of objects, image classification, object detection and semantic segmentation.

In order to solve the problem of object localization and recognition and thus integrate AR into the educational application, we implemented a CNN in Python using *Tensorflow, Keras, Tensorflow Object Detection API* and *OpenCV*. Each neuron of the network performs convolution or cross correlation operation for its receptive field as an output value. Pooling layer of the CNN reduces the spatial size of feature maps obtained from the previous layer without changing the number of maps; each neuron "compresses" its receptive field by performing *Max Pooling* function (taking maximum over the receptive layer). Pooling layer reduces sensitivity to slight shifts of the input image and helps to reduce the size of the following layers.

In the course of the work with the CNN a learning dataset was prepared; object classification and localization in the image were performed successfully with quality of detection improved by using the sliding window technique and non-maximum suppression algorithm. In order to optimize the search of compliances by sliding window technique each pixel of the image of size m x n was estimated by formulae (1-4) [5]

$$M_{m,n} = \sum\nolimits_{\forall(u,v)\in\varepsilon}(C^1_{u,v} + C^2_{u,v} + C^3_{u,v}), \tag{1}$$

where $(u,v)\in\varepsilon$ are conditions checked in every position of the pixel of the template which is centered at (*m,n*). If the conditions are satisfied, $C^k = 1$, otherwise $- C^k = 0$.

$$C^1_{u,v} : (\varepsilon_{u,v} = bg) \cap (F_{u,v} = bg), \tag{2}$$

$$C^2_{u,v} : (\varepsilon_{u,v} = fg) \cap (F_{u,v} = fg) \cap (|d_{u,v} - d_{Hi}| < d_{\max}), \tag{3}$$

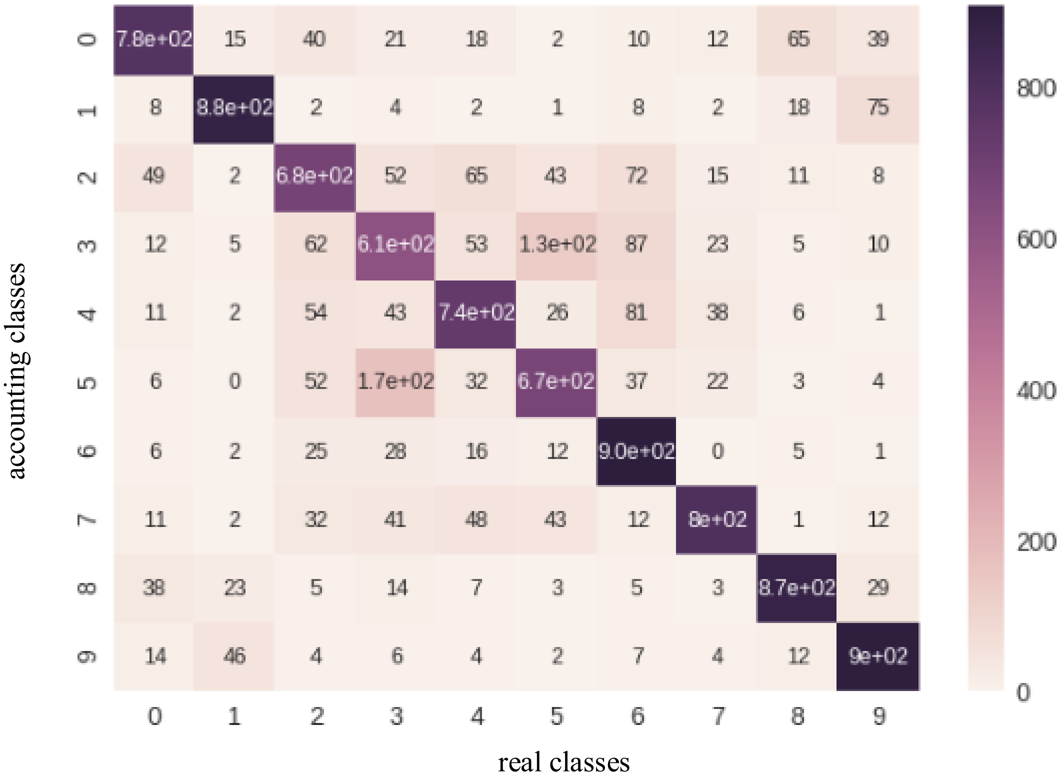
$$C^3_{u,v} : (\varepsilon_{u,v} = bg) \cap (F_{u,v} = fg) \cap (|d_{u,v} - d_{Hi}| < d_{\max}). \tag{4}$$

C^2 and C^3 are conditions providing robustness of estimation against isolated overlaps; bg is background and fg is foreground. Pixel (*m,n*) with maximal estimation $M^H = \max\{M_{m,n}\}$ is taken as the most probable head pose center in the scene.

On the stage of estimating object position each feature's estimation is performed throughout the whole image, which provides numerous assumptions that are then checked by classifiers. After that multiple detections are removed by non-maximal suppression algorithm that handles smoothing, filters out noise by blurring edges, finds intensity gradients, suppresses non-maximals and performs threshold filtration and edge tracking by hysteresis. Since the issue of image classification is basically determined by semantic gap (for a photo is merely a matrix of numbers [0, 255]) we use a confusion matrix based on cross-tabulation to demonstrate the correlation between the values of the complying classes obtained from different sources [6]. Figure 2 presents confusion matrix where all values located on the diagonal demonstrate the compliance between predicted classes and actual data (correct classification). Elements outside the diagonal reflect non-compliance between predicted and actual classes (classification errors), the sum of diagonal values is the number of pixels correctly classified (5):

$$overall\ accuracy = \frac{n_{AA} + n_{BB} + n_{CC} + n_{DD} + n_{EE}}{N} \tag{5}$$

where $n_{AA}, n_{BB}, n_{CC}, n_{DD}, n_{EE}$ are the diagonal elements, N – total of the pixels in the matrix.



	0	1	2	3	4	5	6	7	8	9
0	7.8e+02	15	40	21	18	2	10	12	65	39
1	8	8.8e+02	2	4	2	1	8	2	18	75
2	49	2	6.8e+02	52	65	43	72	15	11	8
3	12	5	62	6.1e+02	53	1.3e+02	87	23	5	10
4	11	2	54	43	7.4e+02	26	81	38	6	1
5	6	0	52	1.7e+02	32	6.7e+02	37	22	3	4
6	6	2	25	28	16	12	9.0e+02	0	5	1
7	11	2	32	41	48	43	12	8e+02	1	12
8	38	23	5	14	7	3	5	3	8.7e+02	29
9	14	46	4	6	4	2	7	4	12	9e+02

Figure 2. Confusion matrix demonstrating correlation between the values of the complying classes.

The confusion matrix reflected the high level of correct object recognition according to the examined classes, which substantiates the choice of the neural network architecture as a basis for educational application development.

Hidden associative layers enable the CNN to form hypotheses based on finding complicated dependences [7] so that the network itself can use the hidden layers situated next to “receptors” to enable their neurons to activate upon detection of straight lines and various angles, then react on corners, squares, circles and primitive patterns such as alternating stripes, geometric grid patterns etc., thus forming hierarchical associations (figure 3).

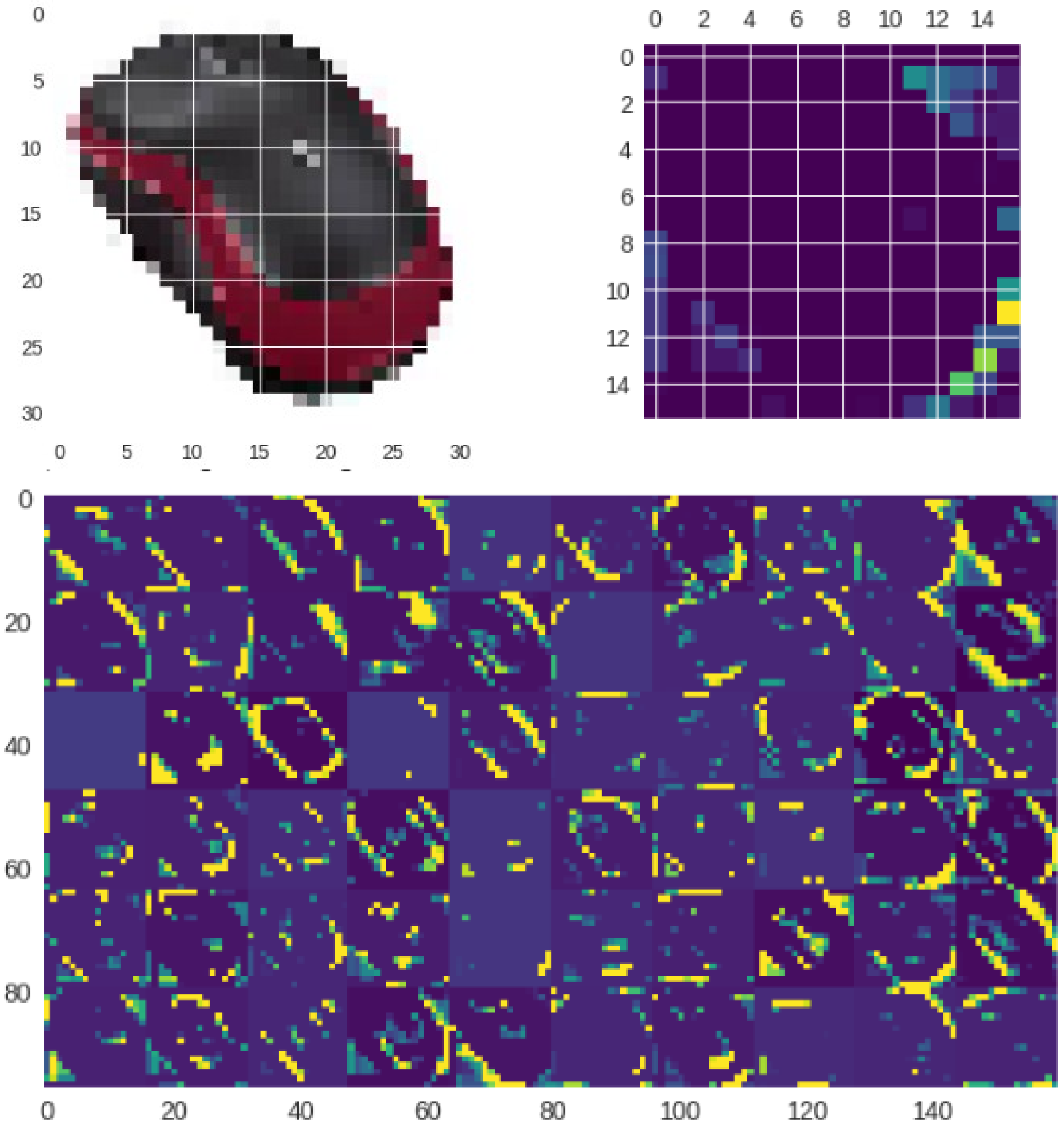


Figure 3. Formation of hierarchical associations in the process of CNN learning.

Before running the educational application the neural network learns using a prepared set of images. Upon finishing the learning the system receives a consequence of input images via the smart phone (webcam). The system then identifies the object in the scene with the help of image detection algorithms in *OpenCV*, analyzes its contour and/or defines if there are any markers present. Then the neural network is applied to perform image recognition based on multiple points to find the affine transformations parameters which suggest algorithms for compression, expansion or distortion of image so that the found key points would comply. Finally, the virtual objects are superimposed on the image. The graph on figure 4 presents quality evaluation of the model learning, based on test data; precision rate approaching 1, the learning is considered to be successful.

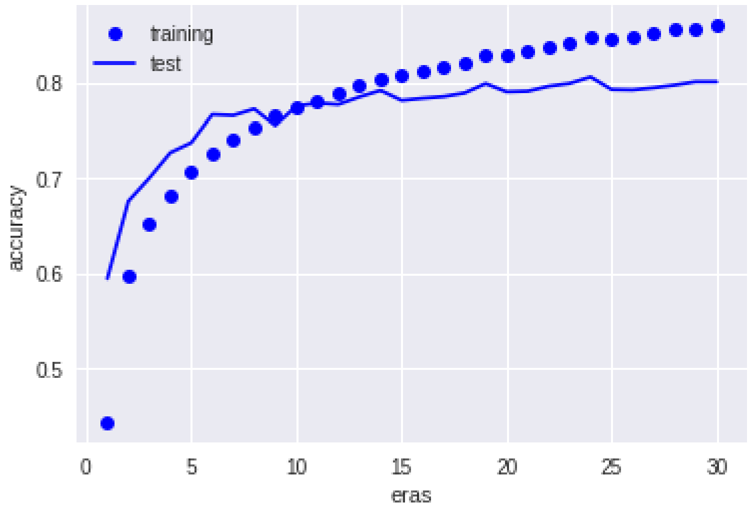


Figure 4. Graph evaluating the quality of model learning.

As the result, neural network learning demonstrates high-quality accuracy of object localization and image recognition. This allows the system to identify the learning material object precisely and superimpose an AR object on it to provide highly informative visualization of educational content.

4. Conclusion

The presented model of AR-powered educational application based on neural network provides opportunities to automate the search for relevant education information eliminating extra data yet unnecessary to process. It can also help to increase the calculating speed of complicated AR algorithms.

AR technologies can be used to provide additional information support and visualization of educational content. The advantages of using AR include a more thorough information support for a studied subject, advanced visual and contextual learning tools, safe training conditions for prospective physical interactions (simulators and virtual laboratory research) and interaction with objects inaccessible in real life.

References

[1] Palmarini R, Erkoyuncu J A, Roy R, Torabmostaedi H 2018 A systematic review of augmented reality applications in maintenance *Robotics and Computer-Integrated Manufacturing* **49** 215-228

[2] Petrov P A, Payor V A,. Panisheva M D 2017 Application of augmented reality technology in aluminum production [in Russian – Primeneniye tekhnologii dopolnennoy real'nosti v proizvodstve alyuminiya] *International Research Journal* **4** 80-82

[3] Soldatov S K, Kuzmina N V 2016 The interface of the future - augmented reality systems [in Russian – Interfeys budushchego – sistemy dopolnennoy real'nosti] *Modern automation technologies* **1** 96–103

[4] Rawat W, Wang Z 2017 Deep convolutional neural networks for image classification: A comprehensive review *Neural computation* **29** (9) 2352-2449

[5] Kingma D P, Ba J L 2014 Adam: A Method for Stochastic Optimization *arXiv preprint arXiv:1412.6980* 15 p

[6] Bekas C, Kokiopoulou E, Saad Y 2007 An estimator for the diagonal of a matrix *Applied numerical mathematics* **57** (11-12) 1214-1229

[7] Dongare A D, Kharde R R, Kachare A D 2012 Introduction to artificial neural network *International Journal of Engineering and Innovative Technology (IJEIT)* **2** (1) 189-194