UDC 004.93'1 004.85

# Capsule Network for Video Segmentation

**Aleksandr Y. Buyko*, Andrey N. Vinogradov*, Igor P. Tishchenko†**

* *Department of Information Technologies*
*Peoples' Friendship University of Russia (RUDN University)*
*6 Miklukho-Maklaya st., Moscow, 117198, Russian Federation*
† *Ailamazyan Program Systems Institute of RAS (PSI RAS)*
*4a Petra-I st., s. Veskovo, Pereslavl district, Yaroslavl region, 152021, Russian Federation*

Email: sas6092@yandex.ru, vinogradov_an@rudn.university, igor.p.tishchenko@gmail.com

In this article, samples of object recognition on video and selection of unique scenes are considered. We used a new algorithm of capsule networks as a tool for video analysis. The algorithm is a continuation of the development of convolutional neural networks. Convolutional networks use a scalar as the base element to be processed. In turn, capsule networks are processing vectors, and use a special routing algorithm. These fundamental differences allow capsule networks to be more invariant to the rotations and changes in illumination of the recognized object. This fact has become the key to choosing this type of networks for analysis of dynamic video. In this article, we propose a method for video segmentation. The essence of this method is as follows. First, you need to determine the main acting objects on adjacent frames. Then, it is necessary to determine whether these objects coincide, if not, then the second frame is considered the moment of transition to another scene. The proposed method was tested on the custom-collected dataset based on videos from YouTube. There were two classes of objects in dataset. The results presented in the article show that we were not able to achieve a high level of accuracy of video segmentation. Also worth noting that the learning process took quite a long time.

**Key words and phrases:** machine learning, video recognition, video analysis, video segmentation, capsule Network, ConvNet.




In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the 1st Workshop (Summer Session) in the framework of the Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Tampere, Finland, 20–23 August, 2018, published at http://ceur-ws.org

## 1. Introduction

The neural networks have already become a powerful tool for solving various non-trivial tasks in the field of computer science. Unfortunately, today there is no universal solution in the field of neural networks and the solution of each case requires special research. Convolutional neural network (CNN) have proven themselves as a relatively effective method for classifying data [1, 2], but such networks have a number of problems and are able to work effectively under certain conditions. The article [3] disturbed the world of neural networks. Geoffrey Hinton has represented a capsule network that could become a new generation of convolutional neural networks. According to the author, the proposed network are to solve the problem of poor translation invariance and lack of information about orientation. It is important to note, Hinton has laid the preconditions for the creation of capsule networks in his previous articles [4, 5]. It is a well-known fact that CNN have problems with turning objects or changing lighting conditions [6, 7]. Moreover, such a problem greatly affects on the effectiveness of the actions recognition on the videos, i.e. on a series of images.

Thus, in this article, we will study the use of a capsule network for pattern recognition on video and the selection of individual scenes. Networks will be aimed at recognizing objects on dynamic frames such. The transition of the scene will be considered in case of a significant change of the frames' characteristics, such as the change of the existing object. Here is the number of articles that inspired us on writing this paper: [8–11].

## 2. Capsule network algorithm

Capsule network in its structure is similar to the convolutional network (Fig. 1a) and includes several layers consisting of capsules (Fig. 1b). Under certain conditions, the capsules can be activated. Each active capsule will be used to select another capsule from the next layer using a special routing algorithm. Routing is a key feature of capsule networks. Capsules are activated depending on various properties of the image, for example, shape, position, color, texture, etc. Thus, each capsule after training will be responsible for some property of objects in the image. It is assumed that the capsule will be active if there is a property on the image for which the capsule responds (Fig. 1c). A vector is obtained at the exit of the capsule. In this work, the vector will represent the probability of belonging to a particular class.

Procedure 1 Routing algorithm.

1: procedure ROUTING$(j|i, r, l)$
2: for all capsule $i$ in layer $l$ and capsule $j$ in layer $(l+1) : bij \leftarrow 0$.
3: for $r$ iterations do
4: for all capsule $i$ in layer $l : \mathbf{c}i \leftarrow softmax(\mathbf{b}i)$ . softmax computes Eq. 3
5: for all capsule $j$ in layer $(l+1) : \mathbf{s}j \leftarrow Picij\hat{\mathbf{u}}j|i$
6: for all capsule $j$ in layer $(l+1) : \mathbf{v}j \leftarrow$ squash$(\mathbf{s}j)$ . squash computes Eq. 1
7: for all capsule $i$ in layer $l$ and capsule $j$ in layer $(l+1) : bij \leftarrow bij + \hat{\mathbf{u}}j|i.\ \mathbf{v}j$
return $\mathbf{v}j$.

The full algorithm of capsule network is represented in original paper [3].

### 2.1. Proposed method (analytics)

To begin with, in this study, we decided to analyze only visual information, which in the vast majority of cases is enough to identify objects on the video. Thus, the image series will be analyzed.

The goal is to design a deep learning architecture to identify the transition points between two different main active objects on the one video. In this paper we aim to use capsule networks instead of neuron ones to construct the deep learning architecture. This kind of networks was chosen as an alternative to CNN, which have obvious drawbacks when analyzing video.

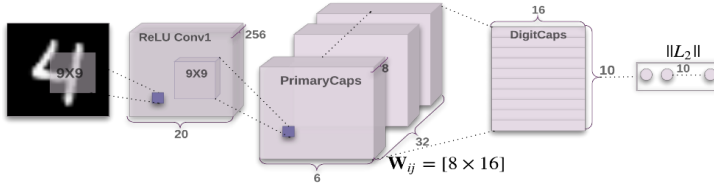


**Figure 1a. The structure of capsule network.**

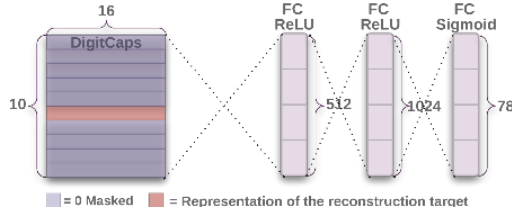


**Figure 1b. The representation of capsule's structure.**

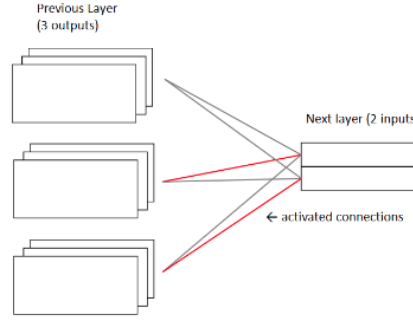


**Figure 1c. Activated interlayer connections (routes).**

First, the CNNs of the scalar and additive nature of neurons in CNNs, neurons on any given layer of the network are ambivalent to the spatial relationships of the neurons within their core of the previous layer and, therefore, within their effective receptive field of given input. In turn, in capsule networks, information on each layer is stored not as a scalar, but as a vector. Such vectors are capable of storing information about such attributes as spatial orientation, magnitude and the other newly-derived attributes depending on the type of capsule layer. The capsules of the upper level are "routed" to the capsule on the next layer using a special dynamic routing algorithm. This algorithm takes into account the consistency between the capsule vectors, eventually forming

meaningful partial relationships that are not found in standard CNNs. The internal representation of the convolutional neural network data does not take into account the spatial hierarchies between simple and complex objects. Therefore, if the image depicts the eyes, nose and lips for a convolutional neural network in an image, this is a clear sign of having a face. A rotation of the object worsens the quality of recognition, whereas the human brain easily solves this problem.

According to the [12], the capsule networks reduce the recognition error of the object in another angle by 45% in comparison with CNN.

The proposed method is described below.

On the input we get an array of images $T[1..n]$. Next, in a cycle, the network receives one image per step. The top preprocessing (compressing) layer is aimed at resize the image to the size of the first convolutional capsule layer, regardless of the size of the original image. For that reason the Convolution algorithm with a Lanczos filter [13] is used.

After exploring several potential architectures we realized that, in fact, to solve the task, there is no need to accurately identify the object to identify the transition points, we decided to use 3 layers of the network to reduce the data abstraction to reduce the complexity of the calculations.

We decided to use the similar model as in [7]. There is summary of three hidden fully connected and output layers of proposed model:

- The first layer process images with the size of $64 \times 64$ pixels.
- The second one is a convolutional layer with a $64 \times 9 \times 9$ sized filters and stride of 1 which leads to 64 feature maps of size $56 \times 56$.
- The third layer is a Primary Capsule layer resulting from $256 \times 9 \times 9$ convolutions with strides of 2. This layer consists of 32 "Component Capsules" with dimension of 8 each of which has feature maps of size $24 \times 24$
- The final output layer includes 2 capsules, referred to as "Class Capsules" which are aimed to return the probability of belonging to a class.

This is how the entire series of images of one video is processed. On the output we get an array with the probabilities of belonging to the objects on the frames to one of the classes in the format $[(p1, p2)1, ..., (p1, p2)n]$. Finally, it is necessary to detect the moment when the main active objects changes. For that case, we simply compare the values of the probabilities obtained. The moment of transition is the case when:

- $p1i < p2(i+1)$
- $p2(i+1) - p1(i+1) < \text{TP}$,

where TP is a transition point probability constant which describes the sensitivity of transitions. The transition constant value is obtained experimentally and depends on the type of video. Obviously, for the image index, when the transition occurred, determine the transition time on the timeline.

Summarizing, this superficial research aims at testing CapsNet as an effective method for analyzing the series of images by detecting the change of the main acting object. Also, we have taken into account [14] while working on this paper.

The general scheme of proposed method is demonstrated on Fig. 2.

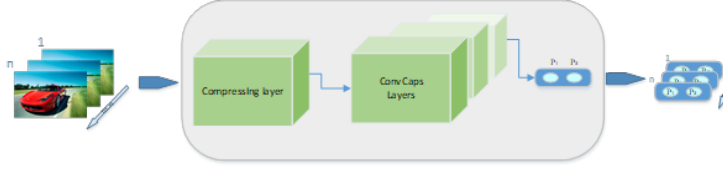


**Figure 2. A simple scheme of proposed method**

### 2.2. Experimental environment

We have prepared a special set of videos, which includes random videos of cars and running people from YouTube with size of $320 \times 240$. The training set included 500 videos randomly cut onto 5s sub-videos and randomly glued together.

Thus, we aimed at recognition of "suddenly-changed" senses. We have used a set of 50 video and a tenfold cross-validation method to assess the accuracy of network.

The method was coded with Python 3.6 and TensorFlow 1.10.0. We used the code template [14, 15] as a start point of ours.

The training process was conducted on a cluster of 4 computers which were build according the results in [16]. Computers had the following characteristics:

- OS: Ubuntu 18.04
- GPU: Nvidia Geforce 540m
- CPU: Intel Core i5|i7
- RAM: 4-8 GB

### 2.3. Results

As a result, we have achieved the best result with an accuracy of 13.31% on a 3 hidden layer network. After a series of experiments it was found that the best results were obtained with TP = 0.4

The results of the assessment test are shown in Table 1.

**Table 1**

**Results of the test:**

| Class | Accuracy, % | Assessment time, s | Average video duration, s | Number of senses |
|---|---|---|---|---|
| No transition | 87.25 | 24 | 21 | 0 |
| People | 7.27 | 13 | 10 | 2 |
| Cars | 9.52 | 14 | 10 | 2 |
| People + Cars | 13.31 | 13 | 10 | 2 |

According to the results, even in the absence of transitions the accuracy level of correctly recognized object was 87.25%. In the case where the video consisted of two scenes with people, the network successfully detected the transition only in 7.27 percent of cases. The transitions on the videos with cars were detected 2.25% better than on video with people. Perhaps this can be explained by video quality and higher dynamics of background change. The results of transitions detection on "People+Cars" videos are illustrated on Fig. 3.

Definitely, we can say that the most of guessed recognized transitions were among different classes of objects. In addition, we assessed the influence of background saturation by its middle value (0 – the darkest, 1 – the brightest). We found out that the background is extremely important in the recognition. The best results have been achieved between the values 0.3 and 0.4.

Given the overall low accuracy of transition detection on 2 scenes videos, we decided not to test the method on the videos with more scenes.

We have to point out the problem with performance. The training lasted about 6 hours, that a bit longer than CNN with similar architecture and parameters.

The drawbacks of such method:

- a small and poorly organized collection of data
- the usage of predetermined number of classes

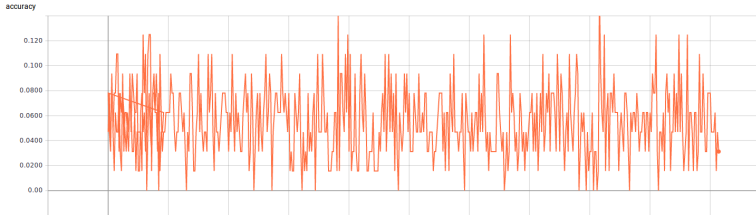


**Figure 3. The results of trained network on "People + Cars" videos.**

– poorly optimized architecture and routings of the network
– weak algorithm for determining object shifts by frames

We are convinced that it is necessary to split the method onto two modules: object recognition and transition detection. The main reason for this breakdown is the need for highly specialized modules capable of solving tasks at different levels of abstraction. However, the integration of these modules will not be obvious. In addition, we suppose, it would be interesting to apply approaches from adjacent areas for data analysis, for example, approaches of dynamic scaling [17] or queuing theory methods [18] to improve the efficiency of calculations.

## 3. Conclusions

To sum up, in this paper we have studded the capsule network in the task of recognizing objects on video and highlighting unique scenes. Despite the stated efficiency of image recognition, capsule networks showed a low level of accuracy of recognition of transitions between scenes on videos.

It is important to note, even in the case of a high level of recognition accuracy the weak point of the CapsNet approach for video recognition is the limitation on classes. People should encode as little as possible the amount of knowledge in the AI software, and instead force them to rely on themselves from scratch. Therefore, currently considered algorithms are not able to provide acceptable image recognition efficiency for the exact segmentation of video. Considering the success of recurrent networks for recognizing actions on video in [19], we can assume that it is worth exploring the proposed in [20] Recurrent Capsule Network.

## Acknowledgments

The publication has been prepared with the support of the "RUDN University Program 5-100". The work is partially supported by state program 0077-2016-0002 «Research and development of machine learning methods for the anomalies detection».

## References

1. A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.
2. P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for video object segmentation. In BMVC, 2017.
3. Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. «Dynamic routing between capsules». Advances in Neural Information Processing Systems 30, pp. 3856–3866, 2017.

4. Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.
5. Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In International Conference on Artificial Neural Networks, pp. 44–51. Springer, 2011.
6. Geoffrey Hinton, Sara Sabour, Nicholas Frosst. Matrix capsules with EM routing (PDF). April 2018.
7. Parnian Afshar, Arash Mohammadi, Konstantinos N. Plataniotis, "Brain Tumor Type Classification via Capsule Networks" / CoRR, abs/1802.10200, 2018.
8. K.-K. Maninis, S. Caelles, Y. Chen, "Video Object Segmentation Without Temporal Information", In CoRR, abs/1709.06031, 2017, http://arxiv.org/abs/1709.06031S.
9. Caelles, K.K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-Shot Video Object Segmentation", CVPR, 2017.
10. G. Bertasius, J. Shi, and L. Torresani. Semantic segmentation with boundary neural fields. In CVPR, 2016.
11. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In CVPR, 2016.
12. Thibault Neveu. Understand and apply CapsNet on traffic sign classification. Becoming Human, November 2017.
13. Wilhelm Burger, Mark J. Burge. Principles of digital image processing: core algorithms. Springer, 2009. P. 231–232. ISBN 978-1-84800-194-7.
14. Max Pechyonkin, "Understanding Hinton's Capsule Networks. Part I: Intuition.", Deep Learning, Nov 3, 2017.
15. Zafar, "Beginner's Guide to Capsule Networks" [https://www.kaggle.com/fizzbuzz/beginner-s-guide-to-capsule-networks] 03.08.2018.
16. Kondratyev A., Tishchenko I. Concept of Distributed Processing System of Image Flow // Robot Intelligence Technology and Applications 4. Results from the 4th International Conference on Robot Intelligence Technology and Applications (RiTA2015) / ed. by J.-H. Kim, F. Karray, J. Jo, P. Sincak, H. Myung. Serie "Advances in Intelligent Systems and Computing", 447 (2016) 479–487. URL: https://link.springer.com/chapter/10.1007%2F978-3-319-31293-4_38. DOI: 10.1007/978.
17. Sopin E.S., Gorbunova A.V. , Gaidamaka Y.V. , Zaripova E.R. Analysis of Cumulative Distribution Function of the Response Time in Cloud Computing Systems with Dynamic Scaling, Automatic Control and Computer Sciences, 52 (1) 2018, Pages 60–66. DOI: 10.3103/S0146411618010066.
18. Gaidamaka Y., Zaripova E., Comparison of polling disciplines when analyzing waiting time for signaling message processing at SIP-server. Communications in Computer and Information Science, 564 (2015) 358–372. DOI: 10.1007/978-3-319-25861-4_30.
19. Buyko AY, Vinogrdov AN "Action Recognition in Videos with Long Short-Term Memory Recurrent Neural Networks", Applied Informatics, 2017.
20. Srikumar Sastry. "Recurrent Capsule Network for Image Generation", 2018.G. Citti, A. Sarti, A cortical based model of perceptual completion in the roto-translation Space, J. Math. Imaging Vis., 24 (2006) 307–326.