УДК 004.838.3
МРНТИ 20.19.21 DOI 10.55452/1998-6688-2021-18-3-75-82

COMPUTER VISION MODEL COMPARISON

NAM D., SAVINA T.
Kazakh-British Technical University, 050000, Almaty, Kazakhstan

Abstract. *The use of machine learning in the medical field is one of the most difficult and thoroughly unsolved problems. Currently, there are many different algorithms for solving problems in the field of diagnostics and segmentation of biomedical images. Researchers are often faced with the challenge of choosing the best method to apply towards their data. We conducted the empirical research and compared 5 algorithms that able to detect anomalies in the medical images: RCNN, Fast-RCNN, Faster-RCNN, Mask R CNN, U-Net, and Residual Neural Network. The advantages of automatic processing of the medical images are apparent: doctors get a convenient software tool that allows them to diagnose the disease faster and reduce possible errors. The task is to study and then select algorithms for further testing on the actual data. The selection and study of algorithms were based on articles describing the architecture and application of computer vision algorithms.*

Keywords: *machine learning, deep learning, neural networks, convolutional neural networks.*

КОМПЬЮТЕРЛЕРДІ КӨРУ ҮЛГІЛЕРІН САЛЫСТЫРУ

НАМ Д., [1]САВИНА Т.
Қазақстан-Британ техникалық университеті, 050000, Алматы, Қазақстан

Аңдатпа. *Медициналық салада машиналық оқытуды қолдану күрделі және мұқият шешілмеген мәселелердің бірі болып табылады. Қазіргі уақытта биомедициналық кескіндерді диагностикалау және сегменттеу саласындағы мәселелерді шешудің көптеген түрлі алгоритмдері бар. Зерттеушілер көбіне олардың мәліметтеріне қолданудың ең тиімді әдісін таңдау мәселесіне тап болады. Біз эмпирикалық зерттеулер жүргіздік және медициналық кескіндердегі ауытқуларды анықтау мәселесін шеше алатын 5 алгоритмді салыстырдық: Fast-RCNN, Faster-RCNN, Mask R CNN, U-Net, R2-Unet және Resediul Neural Network. Автоматты өңдеудің артықшылықтары медициналық кескіндер айқын: ауруды тезірек анықтауға болады, дәрігерлер ыңғайлы бағдарламалық жасақтама алады және деректерді өңдеудегі қателіктердің пайызы азаяды. Тапсырма қойылды – нақты деректер бойынша одан әрі тестілеудің алгоритмдерін оқып, содан кейін таңдау. Алгоритмдерді таңдау мен зерттеу компьютерлік көру алгоритмдерінің архитектурасы мен қолданылуын сипаттайтын мақалаларға негізделген.*

Түйінді сөздер: *машиналық оқыту, терең оқыту, жүйке желілері, конволюциялық жүйке желілер.*

СРАВНЕНИЕ МОДЕЛЕЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ

НАМ Д., САВИНА Т.
Казахстанско-Британский технический университет, 050000, Алматы, Казахстан

Аннотация. *Использование машинного обучения в области медицины является одной из самых сложных и досконально нерешенных задач. В настоящее время существует множество различных алгоритмов для решения задач в области диагностики и сегментации биомедицинских изображений. Исследователи часто сталкиваются с проблемой выбора наилучшего метода, применимого к исследуемым данным. Мы провели эмпирическое исследование и сравнили 5 алгоритмов, которые*

способны решить задачу определения аномалии на медицинских снимках: R-CNN, Fast-RCNN, Faster-RCNN, Mask R CNN, U-Net, и Residual Neural Network. Преимущества автоматической обработки медицинских снимков очевидны: болезнь можно диагностировать быстрее, врачи получают удобный программный инструмент, а также снижается процент ошибок при обработке данных. Была поставлена задача изучить, а в дальнейшем отобрать алгоритмы для дальнейшего тестирования на реальных данных. Отбор и изучение алгоритмов происходили на основе статей, описывающих архитектуру и применение алгоритмов компьютерного зрения.

Ключевые слова: *машинное обучение, глубокое обучение, нейронные сети, сверточные нейронные сети.*

Introduction

The development of Computer vision models dramatically increases with the rise of computing power. If the first convolutional neural networks were useless in the actual cases because of the slow speed and low accuracy, modern state-of-the-art algorithms allow to proceed data in real time for different cases.

In this paper we suggest a comparison of the various contemporary convolution neural networks for solving an instant segmentation task. We analyze six different architectures according to their accuracy, training speed, weaknesses, benefits, growth points and suitability for our future research which is shown on Fig1.

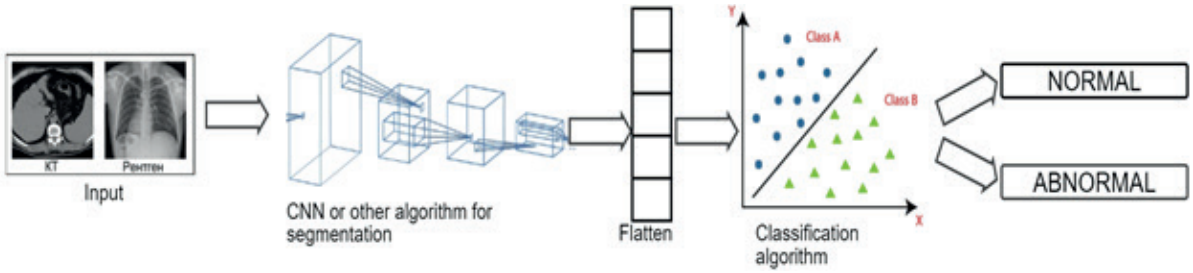

Fig1. Framework architecture of our future work

On the current step we are choosing the appropriate computer vision algorithm for instant segmentation tasks on medical data.

We chose eight algorithms for comparison because they solve similar tasks and are appropriate for biomedical instance segmentation. R-CNN model has a critical disadvantage that makes it inappropriate for the lifetime usage [1]. The algorithm's speed was 13 sec per image on a CPU and 53 sec per image on GPU. The next reviewed article is Faster R-CNN which partly solved the time limitation by updating the loss function and combining RCNN and Spatial Pyramid Pooling NET (SPP Net)[14] algorithms. The next step of development: this branch of CNN also struggled with the same problem, as well as it took into account another key point of computer vision algorithm implementation. It allows to save the disc storage and retrain the model iteratively by the single-stage usage. The speed of Faster R-CNN archived 198ms for proposal and detection both, which have already made it state-of-the-art algorithm. MASK R-CNN is the next observed article, based on Fast and Faster RCNN too. The main difference between MASK R-CNN and Faster R-CNN is the replacement of ROI-Pool with ROI-Align, used for calculating the matrix of features for the candidate region both, but with the bilinear interpolation instead of calculating the matrix of features for the candidate region on borders.

Our main task is mostly based on the medical data. So, we are analyzing appropriate algorithms for the biomedical image segmentation. We found that U-Net, Mask R-CNN, and Res-NET have been already used for medical cases. So, MASK R-CNN and U-net were adopted for Lung Nodules Detection and Segmentation [8] [9], while Res-Net was applied for detection of the diabetic retinopathy [15]. Also, Res-Net can

be used as a backbone for other models to increase their results. The combination of the usage of two or more different architectures in one framework allows the results to dramatically grow up. Thus, R2 U-net integrated the power of U-Net and residual network and allowed the use of historical data. RNN algorithm is based on the LSTM concept and capable of the solving image segmentation tasks. The RNN architecture was used to improve the results of level set-based deformable models (LDM) that are widely used for medical image segmentation by adapting the handcrafted curve evolution velocity. [10]

We determined that models were tested on comparable datasets which allow us to match them by the results given on original articles. According to the results from the original articles we take into account the accuracy score, usually mean average precision was used as a metrics for model evaluation and the speed of image processing on training and testing or both.

Model comparison

R-CNN model shows the high performance mean Average Precision (mAP) of 31.4% [1] on Pascal VOC dataset for the object detection task, but it still has two critical points which have not allowed to use the model in real-time. Because of the complexity of R-CNN architecture, more directly due to the necessity to extract approximately 2k region proposals, it makes the process of training too long for the real-time usage of the algorithm (13s/image on a GPU or 53s/image on a CPU according to the article). Moreover, as it is based on a selective search algorithm it does not allow to retrain the model on this stage. So, the next observed algorithms are struggling with these disadvantages.

Fast R-CNN is the algorithm which has been created by Ross Girshick Microsoft Research. The main goal of this algorithm is to increase the speed of training and testing of existing R-CNN algorithms while saving the accuracy score. Fast RCNN algorithm is based on RCNN and Spatial Pyramid Pooling NET (SPP Net) algorithms which shows the good performance, but it is quite expensive algorithm according to the computation power, because of the complexity.

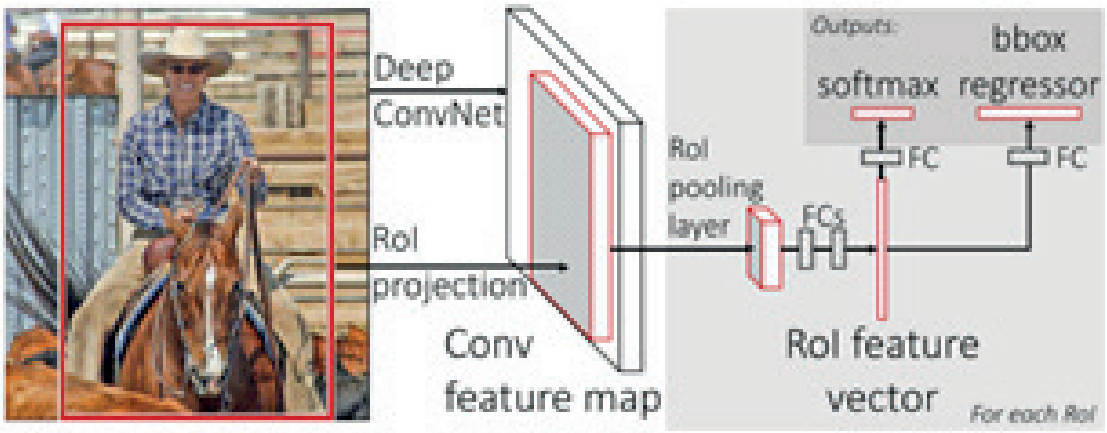

Fig2. Fast R-CNN architecture algorithm. [2]

Basically, the architecture of the model can be described in free steps which are shown in Fig2. The model calculates a conv feature map in the first step using a number of convolution and max-pooling layers. They then used a region of interest (RoI) pooling layer to get the feature vector with a fixed size from the conv feature mask. All vectors are sent to layers that are completely connected. The final two layers are as follows: the first generates softmax probability estimates for K object classes plus a catch-all "context" class (i.e., negative examples for all classes) [1]. For each of the K object groups, the second produces four real-valued numbers.

One more update from the R-CNN algorithm is the multi-task loss for classification and regression both.

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1]L_{loc}(t^u,v)(1),$$

where u,v are target classes, t^u - predicted

tuple $[u \geq 1]$ returns 1 if the expression inside brackets is true and 0 overwise, $L_{cls}(p,u)$= -log p_u is log loss true class u, p - probability. L_{loc}, is defined over a tuple of true bounding-box regression targets for class u, v = (vx, vy, vw, vh), and a predicted tuple tu = (tux , tu y , tu w, tu h), again for class u. [2]

For the evaluation of the model, they used mAP metrics. They tested the model on PASCAL VOC2007, VOC2010 and VOC2012 which contained 20 types of objects. datasets and got the accuracy 70, 68.8m 68.4%.

Faster R-CNN algorithm has positive updates from R-CNN which are caused by the improvement of the architecture of the model:

1. Increasing mAP
2. Single-stage training
3. The usage of multi-task loss
4. Reducing the usage of disk storage

As a consequence, these four advantages allow them to solve problems which were described in RCNN algorithm.

The speed of Fast R-CNN algorithm is near real-time, but it does not take into account the time which is spent on the region proposal. The next algorithm is Faster RCNN which also has been developed by the Microsoft research group.

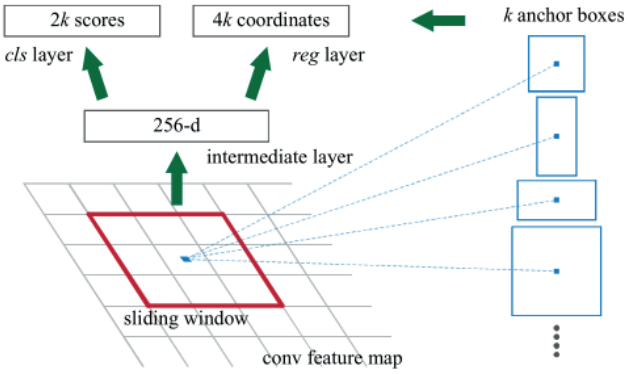

Fig 3. Region Proposal Network (RPN)[3]

The image is used as an input of **Region Proposal Network** (RPN) which is shown in Fig 3, while the output is the set of rectangles with corresponding objectless scores.

The FPN is a fully convolution network with SDG optimizer. They fully inherit the approach of training from the Fast R-CNN algorithm. For mini batch generation they used 256 random anchors with the proportion of positive and negative samples up to 1:1. In case of lack positive samples, they add negative anchors to a mini-batch.

The new layers were created using weights from a zero-mean Gaussian distribution with a standard deviation of 0.01. Image Net creates the rest of the layers. They also tune some ZF and VGG layers to conserve memory. While RPN was used for the region proposed generation, Fast R-CNN was accepted as an algorithm for detection. The dataset from PASCAL VOC 2007 detection championship was used for the evaluation of the model. Overall, it contains 5000 trains and the same test images with 20 types of objects. They used mean Average Precision as the evaluation metrics. The formula is

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

where Q is the number of queries in the set and AveP(q) is the average precision (AP) for a given query, q. (4)

The best result of the model was achieved with the implementation RPN and ZF together. The MAP was 70.4%. According to the speed being one of the factors which has led to creation of this algorithm, 198ms was achieved for both proposal and detection, while the previous algorithm allowed the speed 300ms (0.3 from previous article). While Faster R-CNN is a state-of-the-art algorithm, the creators of **Mask R-CNN** algorithm found that its performance also can be updated in consideration with pixel-to-pixel position of input and output images. The next algorithm is Mask R-CNN [3] which has been presented via Facebook AI research group. It is built on Fast and Faster R-CNN, much as before. [t3. Fig. 4 depicts the architecture of the Mask R-CNN system.

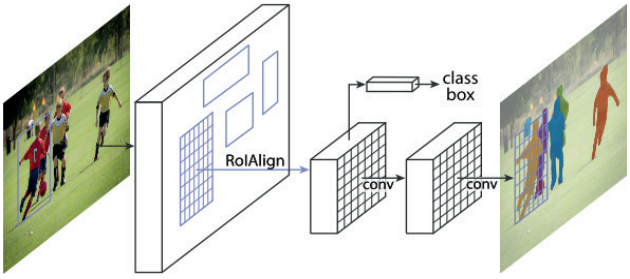

Fig 4.Framework for instance segmentation [5]

Overall, Faster RCNN does not take into account a pixel-to-pixel position of input and output image. Mainly the architecture is separated into two stages: backbone and head for the feature extraction and bounding-box recognition, mask prediction, correspondingly. According to the author of the article, every type of CNN could be adapted as a backbone, while they stopped on Feature Pyramid Network because it allowed to get the efficient speed and accuracy score at the same time. The architecture of the head is shown on Fig 5.

For solving this one ROI pooling layer was replaced via ROI align layer. The difference between them is that the values are rounded to integer in ROI pooling, while in ROI align uses fractional values. As a loss function was used the multi task loss

$$L = L_{cls} + L_{box} + L_{mask},$$

where L_{cls}- classification, L_{box}- bounding box regression,

L_{mask}- mass loss in Mask R CNN. (5)

Also, the significant remark is that MASK R-CNN solved both the problem of instant segmentation and object detection, it also can be used for person segmentation.

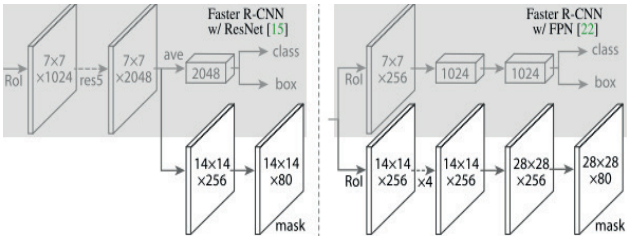

Fig 5. Head architecture

Then this model was adapted for the automatic nuclear segmentation task. [5] They did all experiments on the image set BBBC038v1 from the Broad Bioimage Benchmark Collection [5]. The examples of the images are given on Fig 6. [6]

They achieved maximum Mask Average Intersection over Union 70.54% with ResNet-100 FPN as a backbone on validation data.

The next algorithm is the logical extension of Mask R-CNN. Furthermore, it was the winner of the CACOO challenge of the next year. **Path Aggregation Network for Instance Segmentation algorithm** [7] was the winner on CACO 2017 Challenge instance segmentation and achieved second place in Object Detection task. Overall, CACO dataset consists of 200 000 different images with difficulty to derivation among classes because of blur, number of different objects and other examples of complexity on the image.

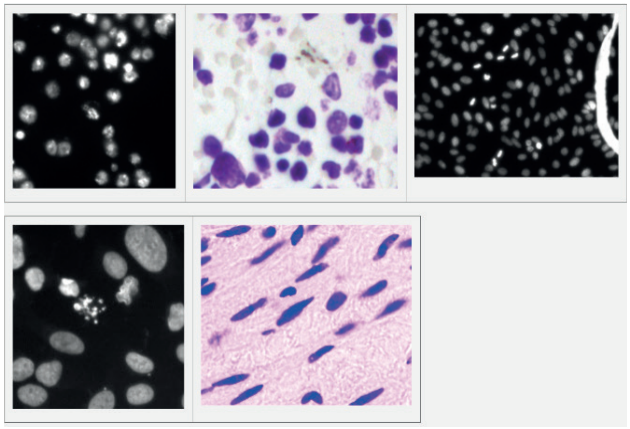

Fig 6. Examples of images [6]

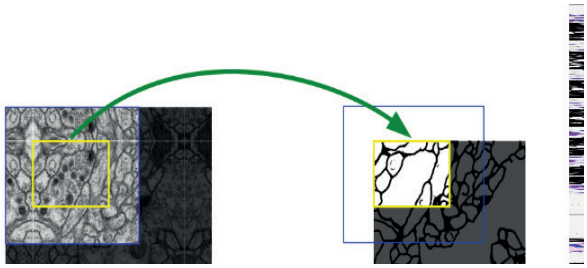

Fig 7. Illustration of framework.

There are basic descriptions of the architecture of PANnet. The first step is Feature Pyramid Network Backbone and Bottom-up path augmentation which were designed for reducing information path. The second step is Adaptive Feature pooling. This part of the Framework can collect feature levels' features for each proposal. And the last one if fully connected network (FCN) from original Mask RCNN with additional properties. The architecture Fig 7.

U-Net belongs to state-of-the-art CNN. It was constructed specifically for biomedical image segmentation [12].

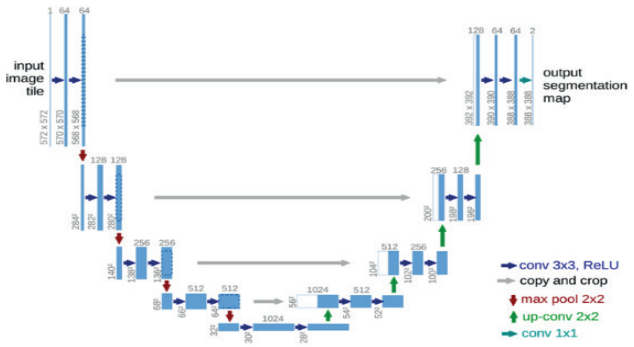

Fig 8. U-net architecture [11]

In the article [11], Unet architecture was described and demonstrated.

The network architecture of Unet is shown in Figure 8. Architecture divided into two hemispheres: on the left side is the contracting path, which following the standard architecture of CNN and on the right side is expanding paths. After combining the images with the data and passing it through other convolution layers, the network design is made up of a series of convolution and pooling layers that minimize the spatial resolution of the picture until increasing it. Generally, the network functions work as a filter. Each compression block takes an input, adds two ReLu 3X3 convolutional layers, and then a pool of maximum compression coefficients.

For each layer in the pool, the number of feature maps is multiplied. The bottleneck layer is made up of two 3*3 Conv layers and a 2*2 Conv layer. Each expansion module sends data to two 3*3 Conv layers and a 2*2 upsampling layer, halving the number of object channels. Also included is concatenation with a correctly clipped object map from the contract direction.

Finally, the 1X1 Conv layer is used to make the output segment count equal to the number of function maps. U-net applies a loss function to each pixel in the image. This makes it easier to spot specific cells in the segmentation diagram. A Softmax value is assigned to each pixel, followed by a loss function. This changes the issue from segmentation to grouping, requiring each pixel to be assigned to one of the classes. The network includes 23 convolutional layer. To ensure a smooth split of the output segmentation map, choose the size of the input tile so that all 2x2 max-pooling operations are applied to a layer with even x and y sizes (see Figure 9).

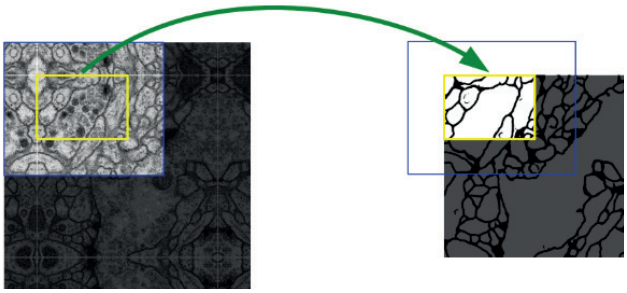

Fig 9. Overlap-tile strategy [11]

Cross entropy is often used as a loss function for UNet:

$$E = \sum_{x \in \Omega} \omega(x) log\ (p_{l(x)}(x)) \qquad (6),$$

where ω - set of multiplicative coefficients, x - pixel position, p - softmax activation function, $l: \Omega \rightarrow \{1,...,K\}$ is the true label of each pixel, K is the number of classes. [1]

The main advantages of Unet:

1. It is a computationally effective method
2. It can be trained with a limited dataset.
3. End-to-end training
4. Preferable for bio-medical applications.

With ResNet-101, the **Residual Neural Network** will replace VGG-16 layers in Faster R-CNN. Many researchers have noticed that this approach has improved. Residual Network (ResNet) is a form of neural network that first appeared in 2015. [13]

In the 2015 ILSVRC classification competition, ResNet won 1st place and entered the top 5 in the COCO 2015 competition for ImageNet Discovery, ImageNet Localization, Coco Discovery and Coco Segmentation. The Resnet neural network is able to effectively train both with 100 layers and with 1000 layers.

ResNet is based on residual learning. In 2015, deep convolutional neural networks were able to classify images better than humans. Previously, many researchers faced such a problem when a deeper network begins to collapse, since with increasing network depth, accuracy first increases and then quickly deteriorates. More layers in conventional neural networks imply a

better network, but due to the disappearing gradient issue, backpropagation will not update the weights of the first layer correctly. As the error gradient propagates back to the earlier layers, re-multiplying makes the gradient small. Thus, as the number of layers in the network increases, its performance saturates and begins to decline rapidly. Res-Net solves this problem with an identification matrix. By using the identity function for backpropagation, the gradient is only multiplied by one. This protects the input and prevents data loss.

The drop-in training accuracy demonstrates that not all networks are simple to improve. To overcome the problem of reduced training accuracy, when optimization is impossible, Microsoft has proposed a deep "residual" training structure. The F (x) + x formulation can be implemented using neural networks with fast access connections (Figure 10).

ResNet uses a skip connection, which means that the original input is also related to the output of the convolution block. This aids in the solution of the gradient fading problem by allowing the gradient to travel along a different direction. They often use an authentication feature that allows the higher tier to perform just as well as the lower tier, if not better.

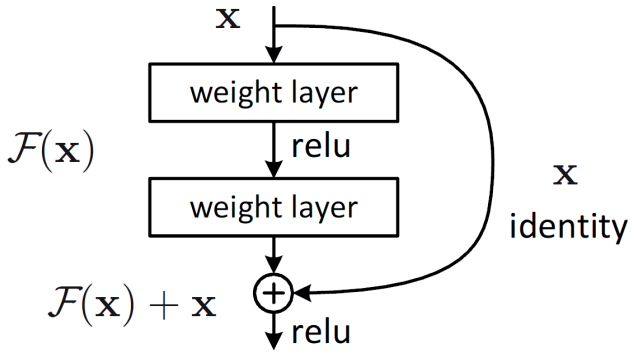

Fig.10. Building block of residual learning. [13]

Resnet's network uses 3*3 filters, stride 2 CNN layers, a global average pooling layer, and a 1000-way fully wired layer with Softmax. Network of ResNet uses a 34-layer simple network architecture inspired by VGG-19, to which a connection shortcut is then added. These fast connections then transform the architecture into a residual network.

The ResNet model, in comparison to VGG networks, has less filters and is less complex. Add a quick link (Figure 11, right) to the simple network mentioned above, which transforms the network into a residual version of the network. When the input and output dimensions are the same, the recognition simple couplings F (x W + x) can be used directly (solid line quick couplings in Figure 11). He considers two choices as the dimensions increase (dotted lines in Figure 11):

To increase the dimension, fast join performs identifier matching with additional zeros added. There are no additional parameters introduced by this option.

Fast connect projection in F (x {W} + x) is used for dimension matching (done with 1 × 1 convolutions). [13][14]

Conclusion

The article reviewed computer vision algorithms that are based on convolutional neural networks. Each subsequent algorithm implements the disadvantages of the previous one. At the moment, the approach of finding important segments using convolutional neural networks is the most popular in computer vision, since algorithms based on transformers require high computing power.

In accordance with the goal of the task: finding the optimal algorithm for image segmentation and further classification of computed tomography images, we chose u-net and Mask R-CNN for further practical testing.

REFERENCES

1. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierar-chies for accurate object detection and semantic segmentation,”Proceed-ings of the IEEE Computer Society Conference on Computer Vision andPattern Recognition, pp. 580–587, 2014.
2. R. Girshick, “Fast R-CNN,”Proceedings of the IEEE International Conference on Computer Vision, vol. 2015 Inter, pp. 1440–1448, 2015

3. S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-timeobject detection with region proposal networks,” 2016.
4. K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, “Mask R-CNN,”IEEETransactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2,pp. 386–397, 2020.
5. J. W. Johnson, “Adapting Mask-RCNN for automatic nucleus segmentation,” arXiv, pp. 1–7, 2018.
6. Image set BBBC038v1, available from the Broad Bioimage Benchmark Collection [Caicedo et al., Nature Methods, 2019]
7. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network forinstance segmentation,” 2018
8. E. Kopelowitz and G. Engelhard, “Lung nodules detection and segmen-tation using 3d mask-rcnn,” 2019
9. C. Zhao, J. Han, Y. Jia, and F. Gou, “Lung nodule detection via3D U-net and contextual convolutional neural network,”Proceedings -2018 International Conference on Networking and Network Applications,NaNA 2018, pp. 356–361, 2019.

A. Pak, A. Ziyaden, K. Tukeshev, A. Jaxylykova, and D. Abdullina,“Comparative analysis of deep learning methods of detection of diabeticretinopathy,”Cogent Engineering, vol. 7, no. 1, p. 1805144, 2020.

10. O. Ronneberger, P. Fischer, T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, vol. 9351, pp.234-241, 2015.
11. M. Zhou, M. Rahman, N. Siddiquee, Tajbakhsh, and J. Liang, Unet++, “A nested u-net architecture for medical image segmentation”, Springer International Publishing, 2018, vol. 11045 LNCS. [Online].
12. K.He, X.Zhang, S.Ren, “Deep residual learning for image recognition”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-decem, p.770-778, 2016.
13. Romera-Paredes and P. H. S. Torr, “Recurrent instance segmentation,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9910 LNCS, pp. 312–329, 2016.
14. O. Vinyals, Q.Le, “A Neural Conversational Model”, vol. 37, 2016.

Information about authors:

1. Nam Diana – Kazakh-British Technical University
Email: d_nam@kbtu.kz
2. Savina Tamara – Kazakh-British Technical University
Email: t_savina@kbtu.kz