Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 54-13 (2021) 588–592

Convolutional Neural Network for Convolution of Aerial Survey Images

Nguyen Van Trong [*] **Pashchenko Fedor Fedorovich** [**]
Le Duc Tiep [***] **Vu Chien Cong** [****]

[*] *Department of Radio Engineering and Cybernetics, Moscow Institute of Physics and Technology, Moscow, Russia, (e-mail: van.chong.nguen@phystech.edu).*
[**] *Intelligent Systems Modeling and Control Laboratory, V. A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia, (e-mail: pif-70@yandex.ru)*
[***] *Department of Control Engineering, Le Quy Don Technical University, Hanoi, Vietnam, (e-mail: letiep1@mail.ru)*
[****] *Operation and servicing of motor vehicles, Moscow Automobile and Road Construction State Technical University, Moscow, Russia, (e-mail: vucongktqd@gmail.com)*

Abstract: The article presents a neural network for convolution of aerial survey images to search and localize objects. When developing a convolutional neural network for convolution of aerial survey images, it is advisable to use the power of cloud technologies, by deploying the CNN on a cloud server. In this article, to construct a convolutional neural network with a full-scale network strategy, we used ResNet, of which architecture is bas. For traditional convolutional functions, neural networks in the process of convolution are characterized by a local receptive field, which can lead to the generation of local features. Encoding long-range contextual information is not performed properly, and the resulting local features can lead to significant potential disagreements between the features under study, which correspond to pixels with the same tags, resulting in inconsistencies within the class. pixels, eventually leading to low recognition efficiency. To solve this problem, the article improved the convolutional neural network for convolution of aerial survey images.

Keywords: Convolutional Neural Network Recognition System, Aerial Survey Images, Convolution, Aerial Imagery, Neural Network.

1. INTRODUCTION

Modern aerial survey depends on technological advances in the field of object recognition and the development of special software. One of the most important tasks is to determine the exact boundaries of objects. The result of the survey is a series of images, which correspond to the selected scanning planes. Based on the obtained images, we can visually determine the presence of object recognition formations and their limits. The problem is that, while the presence of object recognition is sufficiently easy for visual determination due to their characteristic structural features, determining the exact boundaries is a very difficult task, in some cases almost impossible to solve without additional procedures or image convolution. Therefore, there is a need to build a convolutional neural network for convolution of aerial survey images.

2. STATE-OF-ART

Analysis of aerial survey objects strongly depends on choosing the algorithm that is most suitable for its convolution (Sobel (1972); Robinson (1977); Huang et al. (2016); Khaykin (2016); Cheng et al. (2016); Saurav et al. (2019); Ioffe and Szegedy (2015)). In fact, such a task stands before developers of software tools for aerial survey objects scans are equipment, not in front of commercial, which use object recognition methods. When choosing an algorithm, you have to take into account properties of a particular aerial survey image, and features a specific convolution algorithm (Long et al. (2015a); Marmanis et al. (2016); Noh et al. (2015a); Russakovsky et al. (2015a)).

The authors of (Lebedev (2018)) classify convolution methods according to properties, on the basis of which they are performed (similarity of low-level features); image processing strategies (sequential or parallel); image (color or grayscale); if the method used has a built-in (internal) criterion for verifying convolution qualities. In (Huang et al. (2016); Khaykin (2016); Cheng et al. (2016); Saurav et al. (2019); Ioffe and Szegedy (2015); Long et al. (2015a); Marmanis et al. (2016)) convolution methods are divided into three classes, depending on what they are based on: an edge, an area, or pixels. Classification of convolution methods is considered in the paper (Robinson (1977)). These attributes are used to distinguish threshold convolution; morphological convolution and the method of growing areas. In some works (Robinson (1977); Huang

10.1016/j.ifacol.2021.10.513

et al. (2016); Khaykin (2016); Cheng et al. (2016); Saurav et al. (2019); Ioffe and Szegedy (2015); Long et al. (2015a); Marmanis et al. (2016); Noh et al. (2015a); Russakovsky et al. (2015a); Sherrah (2016a); Chen et al. (2018a)), the classification of convolution methods is considered from the point of view of operator participation in the convolution process: interactive, automatic, semi-automatic. To classify convolution methods, most often used are the following task treatment for aerial survey images: threshold methods; edge detection methods; area selection methods; morphological watershed method; Atlas-based methods; clustering methods; artificial neural networks. To evaluate the effectiveness of the application of different method, as a rule, the following indicators are used (Chen et al. (2018b)): sensitivity; specificity; accuracy. Practice shows that the same method can show good results on certain aerial survey images, while ineffective on other images of the same type (Krizhevsky et al. (2012)).

At the same time, there is almost no automatic processing system with convolution of aerial survey images for the search and localization of object recognition. Therefore, this area of research is currently relevant.

In this article, it is necessary to develop a convolutional neural network for convolution of aerial survey images for the search and localization of object recognition with the establishment. The object recognition results show that by selecting a proper set of parameters, a CNN can detect and classify objects with a high level of accuracy and computational efficiency.

3. RESULTS AND DISCUSSION

When developing a convolutional neural network (CNN) for convolution and search for aerial survey objects, it is advisable to use the power of cloud technologies, that is, to deploy CNN on a cloud server, such as GoogleNet, ResNet, QuocNet, or similar services. This will enable shared access (Long et al. (2015b)).

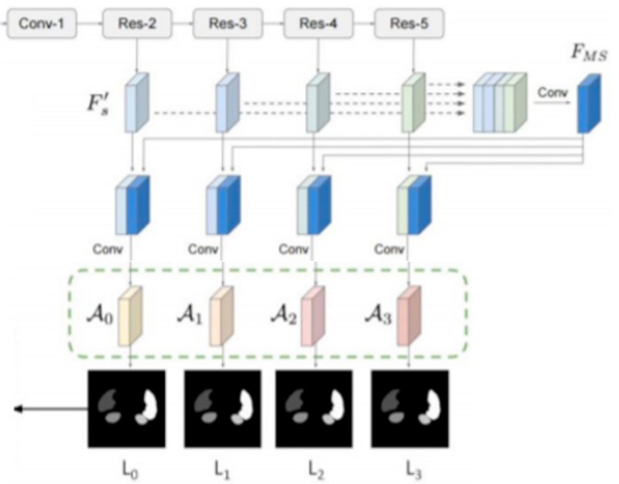
To build an CNN with a full-scale strategy, we will use ResNet. This architecture is the basic one. To improve existing CNN for convolution we will present a general overview of images type of proposed structure (Figure 1).

The multi-scale features are useful in computer recognition problems even before applying deep learning. In the context of deep convolution networks, the integration of multi-scale functions demonstrates amazing performance, which allows us to perform different functions at different scales that helps encode both global and local context.

In this parameter, elements in multiple scales are denoted by F_s, where s specifies the level in the neural network architecture (NM). Because objects have a different resolution for each level s, they are sampled with increasing frequency to the General resolution by linear interpolation, which causes the object blocks to increase to $F_s^{'}$.

Then $F_s^{'}$ of all scales are combined to form a tensor that will be collapsed to create a common bagatoscale system of objects:

$$F_{MS} = conv([F_0^{'}, F_1^{'}, F_2^{'}, F_3^{'}]). \tag{1}$$



Conv - incoming image
Res - criteria for assigning to a layer
$F_s^{'}$*- a specific image block*
Conv - data filter
$\mathcal{A}_i$ *- layers*
L_n *- image layers with identification*

Fig. 1. General type of proposed structure ONR

Such a new structure nm objects are combined from each block of objects at different scales and fed to the control module to create functions for selecting an image area in the form:

$$\mathcal{A}_s = AttMod_s(conv[F_s^{'}, F_{MS}])), \tag{2}$$

where $AttMod_s$ is a control module that views my control area.

Receptive fields in traditional deep convolution models are reduced to local boundaries. This limits the ability to model broader and larger contextual views. On the other hand, channel blocks can be considered as class-specific responses, where different semantic responses are related to each other. Thus, another strategy to improve the representation of specific semantics is to improve the dependencies between channel blocks.

Consider the module for determining the position of the selected image area. Let's denote the input block of objects for the module research area of the image by $F \in \mathbb{R}^{C \times W \times H}$, where C, W, H represent the dimensions, width, and height of the channel, respectively (Figure 2).

Upper branch F passed through a convolutional block, resulting in a feature block: $F_0^p \in \mathbb{R}^{C' \times W \times H}$, where $C' = C/8$. Then F_0^p turns into a block of form attributes $(W \times H) \times C'$.

In the second branch of the block $Finput$ input objects repeat the same operations and will then be transposed, which leads to $F_1^p \in \mathbb{R}^{C' \times (W \times H)}$.

Both blocks are multiplied, and the resulting matrix is overlaid with the maximum layer for generating a spatial block of the selected image area in the form: $S^p \in \mathbb{R}^{(W \times H) \times (W \times H)}$ and it turns out:

$$s_{i,j}^p = \frac{exp(F_{0,i}^p \cdot F_{1,j}^p)}{\sum_{i=1}^{W \times H} exp(F_{0,i}^p \cdot F_{1,j}^p)} \tag{3}$$

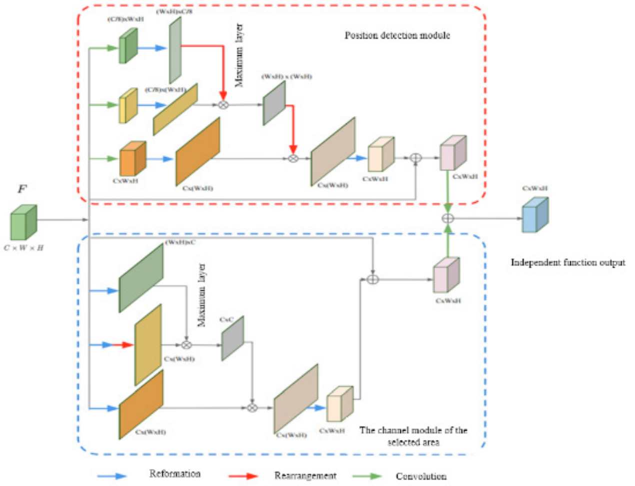


Fig. 2. Modules for determining the position and channel of the selected image area

where $s^p_{i,j}$ evaluates the impact the i-th position have on the j-th position.

Entrance F fed to another convolutional block of the third branch, resulting in $F^p_2 \in \mathbb{R}^{C\times(W\times H)}$, which has the same form as F. Just like in other branches, F^p_2 turns into $F^p_2 \in \mathbb{R}^{C\times(W\times H)}$, and then it is multiplied by permuted ad block version S show area of the image, the output of which turns into: $\mathbb{R}^{C\times(W\times H)}$.

Image area functions, that they correspond to the position detection module, i.e. F_{PAM}. Therefore, it can be formulated as follows:

$$F_{PAM,j} = \lambda_p \sum_{i=1}^{W\times H} s^p_{i,j} F^p_{2,j} + F_j. \tag{4}$$

Meaning λ_p initialized at 0, and gradually taught to render a larger spatial block value to the selected image area. Thus, the position detection module selectively aggregates the global context of the studied features, guided by the spatial block view of selected image area.

Consider the channel detection module (CDM) for the view of selected image area. Entrance $F \in \mathbb{R}^{C\times(W\times H)}$ changed in the first two branches of the CDM and rearranged to the second branch, which leads to $F^c_0 \in \mathbb{R}^{(W\times H)\times C}$ and $F^c_1 \in \mathbb{R}^{C\times(W\times H)}$, respectively.

Then matrix multiplication is performed between F^c and F^c_1 and we get the channel definition block $S^c \in R^{C\times C}$ as follows:

$$s^c_{i,j} = \frac{\exp(F^c_{0,i} \cdot F^c_{1,j})}{\sum_{i=1}^{C} \exp(F^c_{0,i} \cdot F^c_{1,j})}, \tag{5}$$

where $F^c_{i,j}$ is the impact i-th channel on j-th and is defined as $s^c_{i,j}$.

Then multiplication is performed by the transposed version of the input F, that is F^c_2, the result of which turns into $R^{C\times(W\times H)}$. At the end of both view modules for my image area, the newly generated features are passed to the convolutional layer before performing an element-wise summation operation to generate features in the positional channel.

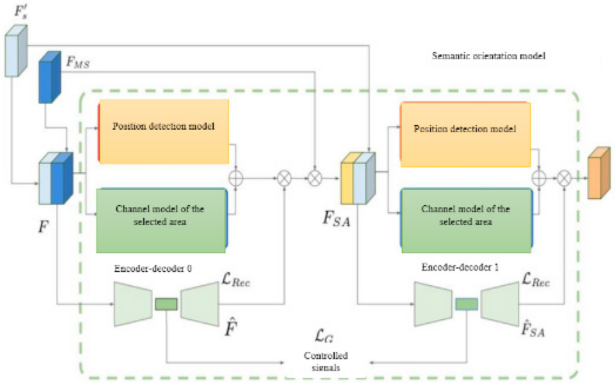


Fig. 3. Module of the selected image area with semantic orientation for the scale s

Similar to CDM, the final block for determining the channel of cosmic snapshots is obtained as (Noh et al. (2015b)):

$$F_{CDM,j} = \lambda_p \sum_{i=1}^{W\times H} s^p_{i,j} F^p_{2,j} + F_j. \tag{6}$$

where IR controls the importance of the channel definition block over the input block of IR characteristics.

Similar to λ_P, λ_c must first be set to 0, and will gradually learn. This formulation combines weighted versions of functions of all channels into output functions, highlighting class-dependent blocks of classes and increasing the differentiation of objects between classes. At the end of both modules of the selected image area, newly generated features are passed to the convolutional level before performing an element-by-Element summation operation to generate features in the positional channel.

Module of the selected image area takes into account the module of characteristics F at the input of the control module of the function of the selected image area at scale. Generated by combining F_{CDM} and Fs', this module in turn the functions of the selected image area using multi-stage refinement (Russakovsky et al. (2015b)).

A general view of the module of the selected image area with semantic orientation for the s scale is shown in Figure 3.

At the first stage, F is used by position and Channel detection modules to generate self-monitoring functions. In parallel, it is necessary to integrate an encoder-decoder network that compresses input functions F in a compressed view and in Hidden Space. The problem is that class information can be embedded in the second position channel module, so that the semantic representation of both encoders and decoders is close, which can be formulated as:

$$L_G = ||E_1(F) - E_2(F_{SEG})||^2_2 , \tag{7}$$

where $E_1(F)$ and $E_2(F_{SEG})$ are encoded representations of the first and second encoder-decoder networks, respectively, and F_{SEG} these are the characteristics generated

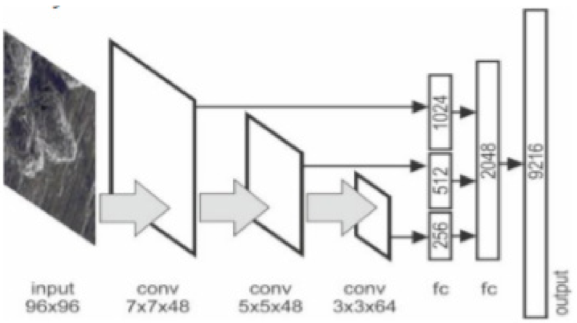


Fig. 4. Architecture of the developed neural network for semantic segmentation of surface images

after the first dual processing module. In particular, feature blocks restored in the first encoder-decoder ($n = 0$) – are then combined with independent features generated by the first module using the matrix multiplication operation to generate F_{SEG}.

In addition, to ensure that the recovered elements correspond to the input functions of the position channel detection modules, the output signals of the encoders must be close to their inputs:

$$L_{Rec} = ||F - \hat{F}||_2^2 + ||F_{SEG} - \hat{F}_{SEG}||_2^2, \tag{8}$$

where $\hat{F}$ and $\hat{F}_{SEG}$ – restored blocks of characteristics, i.e. $\mathbb{D}_0(\mathbb{E}_0(F))$ and $\mathbb{D}_1(\mathbb{E}_1(F_{SEG}))$, the first and second encoder-decoder networks.

Since the modulus of the selected image area is applied at multiple scales, the combined controlled losses for all modules will have the following ratio (Sherrah (2016b)):

$$L_{G_{Total}} = \sum_{s=0}^{S} L_G^s \tag{9}$$

Similarly, total recovery losses become equal:

$$L_{Rec_{Total}} = \sum_{s=0}^{S} L_{Rec}^s \tag{10}$$

where L_{Rec_1} and L_{Rec_2} - recovery losses for encoder-decoder architectures in the first and second module blocks of the selected image area.

Image area module, what stands out, taking into account the characteristics module F at the input of the control module image area functions, what stands out in scale s produced by coupling F_{Who} and $F_s^{'}$, this module generates view functions for my image area using multi -step refinement.

Similarly, total recovery losses become equal.

We selected 512 pixels format, the input layer of the convolutional neural network will contain 262144 (512 × 512) neurons.

We see that the proposed method even surpasses the results of SSD960 and again this is because the proposed method exploits higher resolution (the original high resolution of proposals in original images) and multi-action information.

Results dataset show that the proposed method is more accurate compared to using only CNN for aerial action detection.

4. CONCLUSIONS

In the article, a neural network for convolution of aerial survey images for the search and localization of object recognition was describe. When developing a convolutional neural network for convolution of aerial survey images, it is advisable to also employ cloud technologies, such that the convolutional neural network can be deployed on a cloud server, like GoogleNet, ResNet, QuocNet, etc. The article also demonstrates how to construct and use a convolutional neural network with a full-scale network strategy. While traditional convolutional neural networks, characterized by a local receptive field in the convolution process, can lead to the generation of local features, encoding long-range contextual information will not be performed properly. Consequently, the resulting local features can lead to significant potential disagreements between the features under study, which correspond to pixels with the same tags, hence, resulting in inconsistencies within the class. pixels, and may eventually lead to low-efficiency speech recognition. To solve this problem, the article builds associations between functions and improves the convolutional neural network for convolution of aerial survey images.

Based on the developed convolutional neural network for convolution of aerial survey images can establish a more exactly.

REFERENCES

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X., and Wei, X. (2018a). Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(2), 173–177.

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X., and Wei, X. (2018b). Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(2), 173–177.

Cheng, G., Zhou, P., and Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12), 7405–7415.

Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., and Pan, C. (2016). Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1835–1838. IEEE.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. PMLR.

Table 1. Results of different method for action detection

Method	mAP@0.5
SSD512 (Huang et al. (2016))	15.39%
SSD960 (Huang et al. (2016))	18.80%
CNN	28.30%

Khaykin, S. (2016). Neural networks: full course [neural networks: a complete course]. *Moscow: Williams Publ.*

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 1097–1105.

Lebedev, V. (2018). Acceleration algorithms forconvolutional neuralnetworks. Technical report, Mosk. fiz.-tehn. inst (state university), Dolgoprudny.

Long, J., Shelhamer, E., and Darrell, T. (2015a). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Long, J., Shelhamer, E., and Darrell, T. (2015b). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016). Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, 3, 473–480.

Noh, H., Hong, S., and Han, B. (2015a). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1520–1528.

Noh, H., Hong, S., and Han, B. (2015b). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1520–1528.

Robinson, G.S. (1977). Edge detection by compass gradient masks. *Computer graphics and image processing*, 6(5), 492–501.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015a). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015b). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

Saurav, S., Gidde, P., Singh, S., and Saini, R. (2019). Power line segmentation in aerial images using convolutional neural networks. In *International Conference on Pattern Recognition and Machine Intelligence*, 623–632. Springer.

Sherrah, J. (2016a). Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585.*

Sherrah, J. (2016b). Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585.*

Sobel, I. (1972). Camera models and machine perception. Technical report, Computer Science Department, Technion.