Comparative Analysis of Deep Neural Networks Architectures for Visual Recognition in the Autonomous Transport Systems

Y S Ivanov¹, S V Zhiganov², N N Liubushkina¹

¹Komsomolsk-na-Amure state university, Lenin Avenue 27, Komsomolsk-on-Amur, 681013, Russia
²Deco Systems, Lenin Avenue 30, Kaliningrad, 236022, Russia

Abstract. This paper analyses and presents an experimental investigation of the efficiency of modern models for object recognition in computer vision systems of robotic complexes. In this article, the applicability of transformers for experimental classification problems has been investigated. The comparison results are presented taking into account various limitations specific to robotics. Based on the results of the undertaken studies, recommendations on the use of models in the marine vessels classification problem are proposed.

1. Introduction
Obstacle avoidance is one of the most important tasks of autonomous robots. Besides, it is important to choose a collision avoidance route, taking into account the rules and restrictions. In maritime navigation, such rules are COLREG, based on visual identification of obstacles and vessel classes.

Deep neural networks have shown their efficiency in different domains [1, 2, 3]. Several approaches [4, 5], such as those based on deep neural networks [6] were used to recognize vessel classes. In recent years, various SOTA approaches have emerged. Each of which was efficient in terms of speed or accuracy.

For example, the family of MobileNet models [7] due to the Depth-wise Separable Convolution can significantly reduce the requirements for computational resources without significant loss of accuracy.

EfficientNet networks [8] were obtained using a method of automatic architecture creation by combining the depth and width (number of channels) of the network. The use of EfficientNet Lite models allows replacing MobileNet in some tasks.

At the same time, it is promising to use transformers with attention blocks, which have made a real revolution in the natural language processing (NLP). Vision Transformers (ViT) [9] can achieve better results than most modern CNNs on a variety of image recognition datasets while using significantly less computational resources.

However, it should be noted that most tests are carried out on fairly “clean” datasets such as MNIST or CIFAR, and, often, are far from real conditions.

Many research groups are working on the development of algorithms for recognizing surface objects from their images.

The authors [10] propose to classify ships using Inception and ResNet. In other paper [11] adopts data augmentation and fine-tuning to further improve and optimize the baseline VGG16 model. The
The proposed model attains an average classification accuracy of 97.08% compared to the average classification accuracy of 88.54% obtained from the baseline model.

The aim of the work is to carry out a comparative analysis of modern architectures and SOTA models for visual recognition in relation to the problem of marine vessel classifying by the USV camera.

2. Problem statement

It is necessary to train the most advanced models for visual image classification and test them on the metrics of accuracy, speed and model size.

Mathematical formulation of the object recognition problem in the frames of a video stream is given in [12, 13].

Let there be: an image set \( \omega \in \Omega \) prescribed by the features \( x_i, i = 1..n \), the complex of which for image \( \omega \) is represented by vector descriptions \( \Phi(\omega) = (x_1(\omega), x_2(\omega), ..., x_n(\omega)) = x \); a set of classes \( \mathbb{B} = \{\beta_1, ..., \beta_c, ..., \beta_c\} \), \( c \) — a number of classes. A priori information is represented by a training set (dataset) \( \mathbb{D} = \{(x^j, \beta^j)\}, j = 1..L \) given by a table, each line of which \( j \) contains a vector image description \( \Phi(\omega) \) and a class mark \( \beta_k, k = 1..c \). Let us note that the training set characterizes the unknown mapping \( F: \Omega \rightarrow \mathbb{B} \).

According to the available frames \( I_j \) of the video stream \( V = \{I_1, ..., I_r\} \) and a priori information represented by the training set \( \mathbb{D} = \{(x^j, \beta^j)\}, j = 1..L \) for deep learning of the neural network with a teacher it is required to solve the pattern recognition problem: to detect the patterns \( \omega \) in the form of features assessment \( \tilde{x} \) by applying the neural networks realizing the mapping \( F: I_i \rightarrow \tilde{x} \), and classify them using the mapping \( *F: \tilde{x} \rightarrow \beta_k, k = 1..c \) according to the given criterion \( P(\tilde{x}) \) minimizing the classification error probability.

For training and testing the classifiers, the [14] dataset containing 40552 images of 22 vessel classes was assembled.

3. Problem solution

3.1. Exploratory Data Analysis

Basic Exploratory Data Analysis (EDA) was carried out for this dataset.

Figure 1 shows the distribution of the number of images by classes to identify the class disbalance. Note that the classes are partially balanced and each class contains about 2000 images. Thus, we can perform the preprocessing steps and build the models we will use in this comparison.

Figure 2 shows the samples of each class.

Using a pretrained VGG19 neural network, the features of each image were extracted and a subsequent clustering was carried out.

Cross-validation was used to select the optimal number of clusters. The experiments used a range of values from 25 to 250 with a step of 10.
In Figure 3 and Figure4, clusters are shown along the Ox axis, and classes along Oy. Figure 3 shows clustering into 50 clusters, and figure 4 into 25 clusters. Even when choosing the number of the number of clusters close to the number of classes, we see a high interclass correlation.

3.2. Model training for vessel image classifying

When developing an algorithm for surface objects recognition, we have studied various approaches to the algorithm design. The following models were chosen:

- MobileNetV2
- MobileNetV3
- GhostNet
- NASNetMobile
- NASNetLarge
- EfficientNetB0
- EfficientNetB3
- EfficientNetB7
- ViT

To enlarge the dataset, we used augmentation with mirror reflection, rotation and image scaling.

MobileNet and GhostNet architectures are specified in [6, 14]. Using MobileNetV2 we have managed to achieve 96.03% on the “Top 5 Accuracy” metric.

However, recent researches in the neural networks for embedded devices, such as the Nvidia Jetson, have proposed more efficient architectures and approaches. In particular, the GhostNet architecture able
to achieve higher recognition performance (for example, 75.7% top-1 accuracy) than MobileNetV3 with similar computational costs on the ImageNet ILSVRC-2012 classification dataset was proposed. In this model, Ghost blocks are used, which allow calculating a smaller number of feature maps of objects with the same performance.

Using GhostNet, the loss curve reached 0.11, and the “Top 5 Accuracy” reached 97% over 120 epochs. Precision and Accuracy algorithm metrics for the testing sample were 78% and 78.3%.

Further development of the Reinforcement learning (RL) led to the development of the Neural architecture search (NAS) method. NAS has been used to design the networks that match or exceed the performance of the handcrafted architectures.

EfficientNet is a class of new models that resulted from the study of scaling the models and balancing the depth and width (number of channels) of the network, as well as the resolution of images in the network. The authors of [8] propose a new compound scaling method that regularly scales the depth / width / resolution with the fixed proportions between them (Figure 5).

In order to improve the neural network performance, the researchers automatically selected the initial architecture (Figure 6) using AutoML methods.

So EfficientNet-B1 - EfficientNet-B7 were built, with an integer at the end of the name indicating the value of the compound coefficient.

The advantage of using CNN was that they avoided the need to create hand-crafted visuals, instead learning to complete tasks directly from the data “from start to finish.” However, CNN has a number of problems and disadvantages:

- CNN does not take into account the location of features relative to each other
- When you combine a word, information about the exact location of the object in the image is lost.

Most of the models used for pattern recognition tasks are based on convolutional operations. However, the article [9] introduces the use of transformers for image classifying. While CNN uses pixel arrays, ViT divides the image into visual markers. The input sequence consists of a flattened vector (2D to 1D) of the fragment's pixel values. Each fragment is treated like words in a sentence in the NLP problem. The classification is done using MLP as the last layer. The figure 7 shows a complete ViT scheme from the authors’ article [9].
We have chosen 6x6 as a block size.
Figure 8 shows the visualization of splitting the image into fragments.

Figure 9 shows the VIT architecture for the ship classification problem. The training took more than 19 aces with the achieved metric TOP 5 Accuracy 97.29%.

| Layer (type)          | Output Shape     | Param # | Connected to                  |
|-----------------------|------------------|---------|-------------------------------|
| Input_9 (InputLayer)  | [(None, 224, 224, 3)] | 0       |                               |
| data_augmentation (Sequential) | (None, 224, 224, 3) | 7       | input_9[0][0]                 |
| patches_16 (Patches)  | (None, None, 768)  | 0       | data_augmentation[0][0]       |
| patch_encoder_0 (PatchEncoder) | (None, 196, 64)  | 61760   | patches_16[0][0]             |
| layer_normalization_137 (LayerN) | (None, 196, 64) | 128     | patch_encoder_8[0][0]        |
| multi_head_attention_64 (MultiH) | (None, 196, 64) | 66368   | layer_normalization_137[0][0] |
| add_128 (Add)         | (None, 196, 64)  | 0       | multi_head_attention_64[0][0] |
|                       |                  |         | patch_encoder_8[0][0]        |
| layer_normalization_138 (LayerN) | (None, 196, 64) | 128     | add_128[0][0]                |
| dense_167 (Dense)     | (None, 196, 128) | 8320    | layer_normalization_138[0][0] |

Figure 9. Vision Transformers Architecture for vessel classification.
3.3. Results and Discussion
Using a testing sample obtained from the dataset, an experiment was carried out on the equipment with the following parameters: Intel Core i7-5820K CPU, 1080 Ti GPU. Comparison of efficiency of the described neural networks in terms of standard metrics of classification quality is carried out.

The testing sample size is 225711 class-balanced images. The following results were obtained (Table 1) for the metrics: Precision, Recall F1-score by classes and for Accuracy, AUC – in general, and the image processing speed in seconds.

| Model      | Metrics       | Acc. | Time |
|------------|---------------|------|------|
|            | Metrics       |      |      |
| MobileNet  |               | 74.4 | 0.142|
| GhostNet   |               | 78.3 | 0.17 |
| NASNet     |               | 73.4 | 0.7  |
| EfficientNetB0 |           | 71.1 | 0.14 |
| EfficientNetB7 |           | 79.9 | 0.14 |
| VIT        |               | 77.3 | 0.18 |

4. Conclusion
In this article, the applicability of generally accepted neural network architectures in the field of object recognition for robotic systems was investigated.

The application of various architectures of deep neural networks on modern embedded graphics accelerators for detection and recognition of surface objects is proposed.

It is shown that the use of VIT allows one to achieve less waste of computing resources, which is especially important in embedded computing devices.

The simulation of the developed algorithms based on the hardware prototype of the intelligent vessel control system was carried out. The use of modern SOTA architectures allows improving accuracy and performance.

It is shown that networks with additional memory and attention mechanism are superior to classical approaches, which proves the applicability of these architectures to the classification problem.

In future work, it is necessary to investigate the influence of noise and various weather conditions on the quality of recognition.

5. References
[1] Amosov O S, Amosova S G, Ivanov Y S, Zhiganov S V 2021 Using the deep neural networks for normal and abnormal situation recognition in the automatic access monitoring and control system of vehicles Neural Comput Appl. 33(8) pp 3069–83 https://doi.org/10.1007/s00521-020-05170-5
[2] Amosov O S, Ivanov Y S, Zhiganov S V 2017 Human localization in video frames using a growing neural gas algorithm and fuzzy inference Comput Opt. 41(1) pp 46–58 https://doi.org/10.18287/2412-6179-2017-41-1-46-58.
[3] Amosov O S, Ivanov Y S, Amosova S G 2019 Recognition of abnormal traffic using deep neural networks and fuzzy logic In: 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon) IEEE https://doi.org/10.1109/fareastcon.2019.8934327
[4] Barve K G, Lokhande S S 2017 Obstacle detection from unmanned surface vehicle https://www.ijareeie.com/upload/2017/june/67_OBSTACLE__IEEE.pdf.
[5] Xu Q, Zhang C, Zhang L 2017 Deep convolutional neural network based unmanned surface vehicle maneuvering In: 2017 Chinese Automation Congress (CAC) IEEE
[6] Ivanov Y S, Zhiganov S V, Ivanova T I 2021 Intelligent deep neuro-fuzzy system of abnormal situation recognition for transport systems In: Current Problems and Ways of Industry Development: Equipment and Technologies Cham: Springer International Publishing pp 224–33 https://doi.org/10.1007/978-3-030-69421-0_24

[7] Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T et al. 2017 MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications http://arxiv.org/abs/1704.04861

[8] Tan M, Le Q V 2019 EfficientNet: Rethinking model scaling for convolutional Neural Networks http://arxiv.org/abs/1905.11946

[9] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al 2020 An image is worth 16x16 words: Transformers for image recognition at scale http://arxiv.org/abs/2010.11929

[10] Leclerc M, Tharmarasa R, Florea M C, Boury-Brisset A-C, Kirubarajan T, Duclos-Hindie N 2018 Ship classification using deep learning techniques for maritime target tracking In: 2018 21st International Conference on Information Fusion (FUSION) IEEE https://doi.org/10.23919/icif.2018.8455679

[11] Mishra N K, Kumar A, Choudhury K 2021 Deep convolutional neural network based ship images classification Def Sci J. 71(2) 200–8 https://doi.org/10.14429/dsj.71.16236

[12] Pashchenko F F, Amosov O S, Amosova S G, Ivanov Y S, Zhiganov S V 2019 Deep neural network method of recognizing the critical situations for transport systems by video images Procedia Comput Sci. 151 pp 675–82 https://doi.org/10.1016/j.procs.2019.04.090

[13] Amosov O S, Amosova S G, Zhiganov S V, Ivanov Y S, Pashchenko F F 2020 Computational method for recognizing situations and objects in the frames of a continuous video stream using deep neural networks for access control systems J Comput Syst Sci Int. 59(5) pp 712–27 https://doi.org/10.1134/s1064230720050020

[14] Ivanov Y S, Zhiganov S V, Ivanova T I 2019 Intelligent deep neuro-fuzzy system recognition of abnormal situations for unmanned surface vehicles In: 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon) IEEE