Prohibited Item Detection in Airport X-Ray Security Images via Attention Mechanism Based CNN

Maoshu Xu, Haigang Zhang, and Jinfeng Yang(✉)

Tianjin Key Lab for Advanced Signal Processing,
Civil Aviation University of China, Tianjin, China
jfyang@cauc.edu.cn

Abstract. Automation of security inspections is crucial for improving the efficiency and reducing security risks. In this paper, we focus on automatically recognizing and localizing prohibited items in airport X-ray security images. A top-down attention mechanism is applied to enhance a CNN classifier to additionally locate the prohibited items. We introduce a high-level semantic feedback loop to map the targets semantic signal to the input X-ray image space for generating task-specic attention maps. And the attention maps indicate the location and general outline of prohibited items in the input images. Furthermore, to obtain more accurate location information, we combine the lateral inhibition and contrastive attention to suppress noise and non-target interference in attention maps. The experiments on the GDX-ray image dataset have demonstrated the efficiency and stability of the proposed scheme in both single target detection and multi-target detection.

Keywords: Prohibited item · Detection · Attention · CNN

1 Introduction

Airport security is an important guarantee for aviation safety. Prohibited item detection using X-ray screening plays a critical role in defending passengers from the risk of crime, and terrorist attacks [1]. However, during the security screening, uncontrollable human factors always reduce the accuracy and efficiency of inspection [2]. Establishing an efficient and intelligent security inspection system is crucial to promoting the safe operation of civil aviation and ensuring the safety of passengers. The core work of X-ray screening is to distinguish what type of the prohibited items and detect where they are. To achieve automatic security, the computer is required to replace the security inspector to answer the two questions of "What" and "Where". Automatic and intelligent security detection for X-ray images remains an open question. Most of challenges come from the following points: (1) different imaging modes; (2) clutter background; (3) angle variation of the items in imaging; (4) color variation caused by material difference of the items [3–5].

© Springer Nature Switzerland AG 2018
J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 429–439, 2018.
https://doi.org/10.1007/978-3-030-03335-4_37

In recent years, some deep neural networks have achieved remarkable performance in the areas of target recognition and detection. Compared with traditional algorithms, deep learning algorithms have stronger generalization ability and can achieve higher recognition accuracy. Motivated by the convolutional neural network (CNN), [1] has presented a strategy based on deep features to deal X-Ray image recognition problems on a public GDX-ray dataset. A deep multi-layer CNN approach is employed in [6] for the end-to-end entire feature extraction, representation and classification process, which achieves 98.92% detection accuracy. Usually, common CNNs can only complete the task of image classification. For object detection, the more complex the model is, the more labor-intensive supervision information is required. For example, several network architectures have good performance in object recognition and location such as single shot multibox detector (SSD) and faster region-based convolutional neural networks (Faster R-CNN). However, they all require strong supervision information for training, e.g., bounding boxes or segmentation masks. Collecting such a large amount of labeled data is often expensive and time-consuming. Especially for security images, the position of the prohibited items requires professional security personnel to mark. Taking into account this actual situation of security images, these methods cannot be effective in practical applications.

Recently, the work of Cao [7,8], Zhang et al. [9] provide a new idea for intelligent security inspection. Attention mechanism based CNN has achieved great detection performance on nature image set. For target recognition tasks, CNNs have strong anti-jamming and anti-blocking capabilities. For target localization tasks, the attention feedback mechanism can enable the network to achieve the prohibited item localization, while not requiring a strong supervised learning [7]. The attention mechanism can then find out which areas of the image can cause CNN to extract these features that activate the output node. This is very similar to the working mechanism of the human visual cortex: When dealing with these stimuli, we also know where these stimuli come from [8].

In this paper, we apply the model proposed by Cao [7] on the automatic prohibited item detection system to prove that the attention mechanism can also perform well on X-ray image processing. Considering the large amount of noise and interference between prohibited items in the security image, we combined the lateral inhibition [8] and contrastive attention [9] to establish a neuronal stimulus inhibition model. When performing the feed-back propagation, it can effectively suppress noise and interference. Furthermore, to make the algorithm suitable for the security X-ray image set, we optimized the two suppression methods. Finally, the semantic information of the target is mapped to the image space as an attention map, and we can know which areas of the image are most relevant to the target.

The main contributions of this paper are summarized as follows: (1) We introduce the semantic feedback model in CNNs and obtain a cursory target attention map. (2) To cope with noise and interference in the security image, we combine two neural suppression algorithms to establish a neuronal stimulus inhibition model. (3) To improve the practicality of our model, we develop a

multi-target detection strategy. (4) We perform experiments on the GDX-ray dataset, and our method achieves significant performance in both single-target and multi-target detection.

2 High-Level Semantic Feedback Model

When searching for objects, the top-down attention of a person plays the role of regulating neurons in the visual cortex according to the current task and prior knowledge. Same as most attention models [7], we use a CNN to model visual cortex neurons and apply the high-level semantic feedback mechanism on the CNN framework. It can layer-by-layer calculate correlations between each layer neurons and CNN output semantic notes. As shown in Fig. 1, in the feed-forward propagation process of CNN, an X-ray security image of a gun is mapped to one-dimensional semantic space by a CNN classifier, and target category information is obtained. In the feed-back propagation, the semantic information is mapped to image space by semantic feedback model, and the attention map of the gun is shown in the input image.

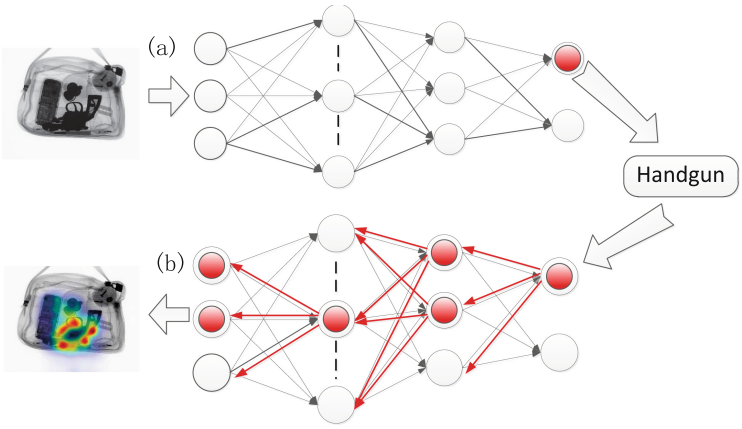


Fig. 1. Attention mechanism based model. (a) and (b) represent feed-forward and feed-back propagation for a convolutional neural network. (a) Given an input image, the output neuron corresponding to the predicted category is activated after the feed-forward propagation and represented by the red dot. (b) In the feed-back propagation, the red dots represent neurons positively related to the output neuron and are activated layer by layer. Finally, we can use the neurons that are activated in the input layer to obtain attention maps. (Color figure online)

The feedback model is much like the backpropogation in the training process. But the signal of the backpropogation changes to the semantic information of the output layer, not the value of loss function. On this basis, the correlation between each convolutional layer and semantic neurons can be calculated by deconvolute

the output with the parameters of the convolutional layer, and is denoted as α^l. This backpropagation is performed from top to bottom as described in Eq. (1), where "$*$" represents deconvolution. In this way, the attention map of the targets is generated which marks the most relevant pixel region in the image to the semantic node, so as to achieve the positioning and segmentation of the targets, as shown in Fig. 2.

$$\alpha^{N-1} = x^N * w^{N-1} \qquad \alpha^{l-1} = \alpha^l * w^{l-1} \quad l = 2, 3, 4, \ldots, N, \tag{1}$$

where x^l denotes the output of the layer, and x^N denotes the output of the network. w^l represents the convolution kernel parameters on layer l.

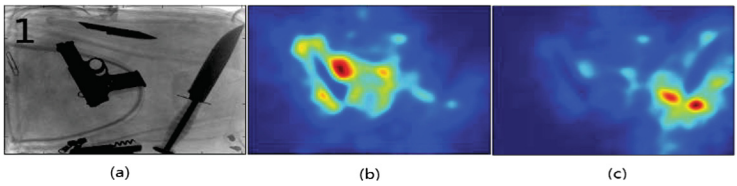


Fig. 2. Result by high-level semantic feedback model for different targets respectively. (a) The input image. (b), (c) Output maps for gun and knife respectively.

3 Neuronal Stimulus Inhibition Model

In the previous section, the attention maps are rather rough, often accompanied by noise and interference. It is because that the X-ray image has a complex background and clutter. Most kinds of noise come from the defect caused by the non-linearity of the network. Since activated patterns can not only be derived from target objects but also derived from background and disturbed objects, there will be a lot of non-target interference in the picture. In order to meet the requirements for precise location of prohibited item, we take the following measures to deal with these noises and interference.

3.1 Lateral Inhibition Mechanism

Lateral inhibition mechanism can enhance the contrasts between the neurons, and has provided good performance on natural images processing [8,10,11]. For further filtering the activated neurons, we apply the lateral inhibition mechanism on the top-down procedure of the CNN frame. Different from [8], in order to deal with a lot of noise in the security X-ray image, we use the output of the previous layer to normalize the suppression coefficient of the current layer. According to the distribution of the value in attention maps, we choose distribution cosine function to evaluation the inhibition values, such as Eqs. (2) and (3).

$$w_{ij}^{ave} = \frac{\cos(\overline{w_{ij}})}{\cos(\overline{x_{ij}})} \tag{2}$$

$$w_{ij}^{dif} = \frac{\sum_{uv} d_{uv} e^{-d_{uv}} \delta(w_{uv} - w_{ij})}{\cos(\overline{x_{ij}})}, \tag{3}$$

$$w_{ij}^{'} = \begin{cases} w_{ij} \; if \quad w_{ij} > (a * w_{ij}^{ave} + b * w_{ij}^{dif}) \\ 0 \quad if \quad w_{ij} < (a * w_{ij}^{ave} + b * w_{ij}^{dif}) \end{cases}, \tag{4}$$

where w_{ij} denotes an element in the normalized attention map of this layer at location (i, j). w_{ij}^{ave} denotes the mean inhibition coefficient, and w_{ij}^{dif} denotes the differential inhibition coefficient. w_{uv} denotes the elements in the sliding window centered on w_{ij}, and d_{uv} denotes the Euclidean distance between w_{uv} and w_{ij}. $\overline{w_{ij}}$ denotes the mean of the elements in the sliding window. x_{ij} denotes the outputs of layer $l-1$. $w_{ij}^{'}$ denotes the new value of the element in the attention map.

Those two kinds of coefficients are standardized by a and b, which we setting $a = 0.2$ and $b = 0.8$. Under the combined effect of these two suppression methods, noise in the attention map can be well suppressed, as shown in Fig. 3. The results prove that the lateral inhibition mechanism has excellent performance in our model.

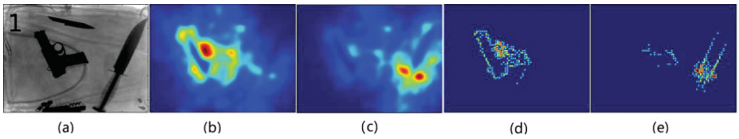


Fig. 3. Comparison between original attention map and the results of lateral inhibition. (a) The input images. (b), (c) Original attention maps for gun and knife respectively. (d), (e) Attention maps after lateral inhibition for gun and knife respectively.

3.2 Contrastive Top-Down Attention

The lateral inhibition model excels in suppressing noise, but when the security image contains multiple prohibited targets, the accuracy of the model is disturbed. In Figs. 3(d)(e), the attention map of a gun is disturbed by the lines around the gun, while the attention map of a knife is interfered by the gun. This is consistent with our previous discussions that other targets in the input image will interfere with the attention map. It has been proved in [9] that using the negation of the original weights of layer can obtain the next level of non-targets signal, and here it is denote as $P_1^{'}$ in Eq. (5).

$$P_1^{'} = EP(-W_1) = A_1 \bigotimes ((-W_1)^{+}(P_0/((-W_1)^{+T} A_1))), \tag{5}$$

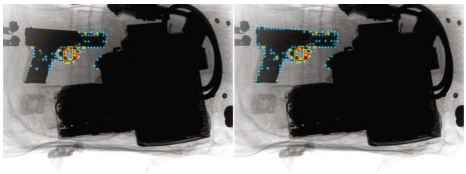
where P_0 represents the probability correlation matrix in the top layer. $P_1^{'}$ represents the probability correlation matrix of the non-targets signal. W^{+} is the

plus weight of the top layer, while $-W$ is the negation. A_1 is the response value of the second layer neurons. $EP()$ represents the function of $-W$ that calculate the non-targets signal.

Because there are too many sundries in the security X-ray images, even in the non-target signal, there are still some target signals. If we immediately subtract it from the semantic signal, the target signal in semantic information will be loss, as shown in Fig. 4(a). Instead of immediately subtracting $P_1^{'}$ from the semantic signal as the [9] did, we use the improved algorithm to obtain more complete semantic information in Eq. (6). In our method, after performing the lateral inhibition on $P_1^{'}$, the target signal in $P_1^{'}$ will be suppressed. When we subtract it from the semantic signal, the target signal can be better preserved, as shown in Fig. 4(b). In this way, we also complete the organic combination of the two algorithms, as shown in Fig. 5. We use contrast attention to remove non-target interference in semantic signal. Furthermore, we add lateral suppression layers in the top-down feedback path to optimize the contrast attention and suppress noise. As shown in Figs. 6(d)(e), it is obvious that compared with the contrastive attention algorithms, the combined algorithms has more powerful ability to suppress noise in the attention map.

$$S = Lat(EP(W_1)) - Lat(EP(-W_1)), \tag{6}$$

where S is the target signal. $Lat()$ is the lateral inhibition function.



(a) (b)

Fig. 4. (a) The result of contrastive top-down attention method. (b) The result of our method

4 Multi-target Detection Strategy

Our model can be used not only for single-target detection but also for multi-target detection. In order to detect every type of threat targets in security X-ray images as much as possible, we designed the following multi-target inspection process:

(1) Given an X-ray image and perform forward to obtain probability values of various types of threat objects.
(2) Judge the objects with a probability value greater than r as the suspected targets, where r is determined in advance.

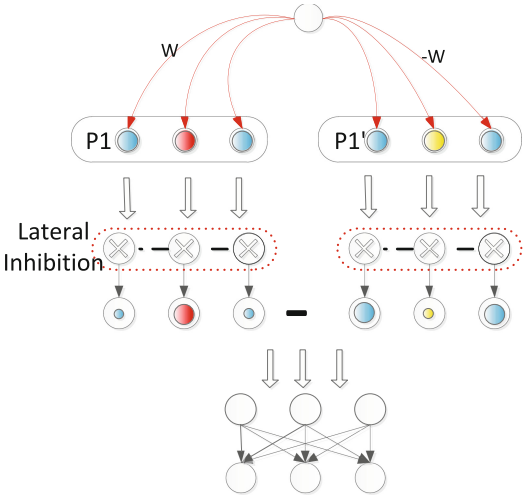


Fig. 5. Neuronal stimulus inhibition model. The red dot represents the target signal and the blue dot represents the non-target signal. The yellow dot represents the target signal in $P_1^{'}$ (Color figure online)

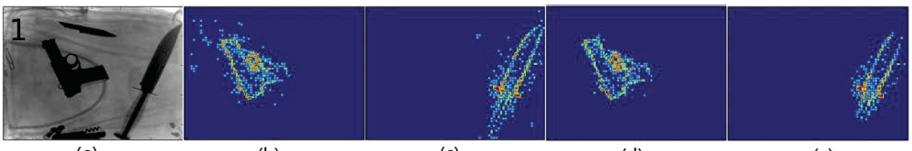


Fig. 6. Comparison between contrastive top-down attention maps and the results of combined algorithm. (a) The input images. (b), (c) Contrastive top-down attention maps for gun and knife respectively. (d), (e) Attention maps obtained by the combination of two algorithms.

(3) Activate the CNN output node corresponding to the target type and perform feed-backward propagation using feedback models.
(4) Obtain response-based attention maps for suspected object at the data layer and display it on the input image.

5 Experiments

To quantitatively demonstrate the effectiveness of the attention mechanism based CNN, we carry out the experiments on the GDX-ray dataset. As mentioned before, our model should answer two questions: what and where. So, we use the above security inspection strategy to verify the model's recognition and location capabilities.

5.1 Dataset

The security image data set used for training and testing in this paper comes from the GDX-ray dataset. The dataset captures single-target and multi-target X-ray security images of guns, knives, darts, and other dangerous goods from multiple angles. However, the original purpose of the database was created to study the traditional computer vision algorithms. If it is used to train the CNN model, there are the following disadvantages: (1) The samples set is very small. There are few single-target images used to train the classification network, and it is easy to cause overfitting; (2) The background is monotonous. Therefore, we performed data augmentation [12] on the GDX-ray dataset. We cut it to 2/3 of the security images original size. The cropped position is random, and we pick out 10 images containing the complete target from the cropped images. Then rotate these 10 images at, 90, 180, and 270° and flip horizontally. Finally one image can be expanded to 80 images. We expanded the data set to 5000 pictures containing 4 object categories. We extracted 90% of the pictures from the dataset as a training set, and another 10% as a testing set.

5.2 CNN Classifiers

Although data augmentation has been carried out, the existing data volume is not enough for the CNN network to fully learn the characteristics of various threat targets. So we use the transfer learning strategy to train the CNN model. The Google net which is pretrained with ImageNet 2012 training set is obtained from Caffe Model Zoo website. We replace the last fully connected layer of the network with a convolutional layer and initialize it. In the training process, the learning rate of the bottom network is set to 0.0001, and the learning rate of the top network is set to 0.001. After 30 iterations with batchsize set of 8, the network tends to convergence.

5.3 Experimental Results

Our model has achieved a classification accuracy of 97.6%. We apply the attention mechanism to single-target and multi-target detection. In order to better coordinate with the security inspector, we directly display the salience map of the target on the input image, 65% positioning is correct after judging by security professionals, as shown in Table 1.

When there is only a single target in the image, even if the gun's pose and the complexity of image background have changed as shown in Figs. 7(a)(b), the target salience map generated by our method can still provide accurate positioning. When there are multiple targets of the same kind in the image as shown in Figs. 7(c)(d), each target is effectively marked in the salience map. When there are multiple types of targets in the image, our model can generate a salience map for each type of target after the multi-target detection process proposed in Sect. 4.

Table 1. Detection accuracy of single-target detection.

Category	Recognition accuracy	Positioning accuracy
Revolver	95.6%	53%
Gun	98.3%	73.5%
Revolver	99.2%	79.3%
Knife	97.2%	54.1%

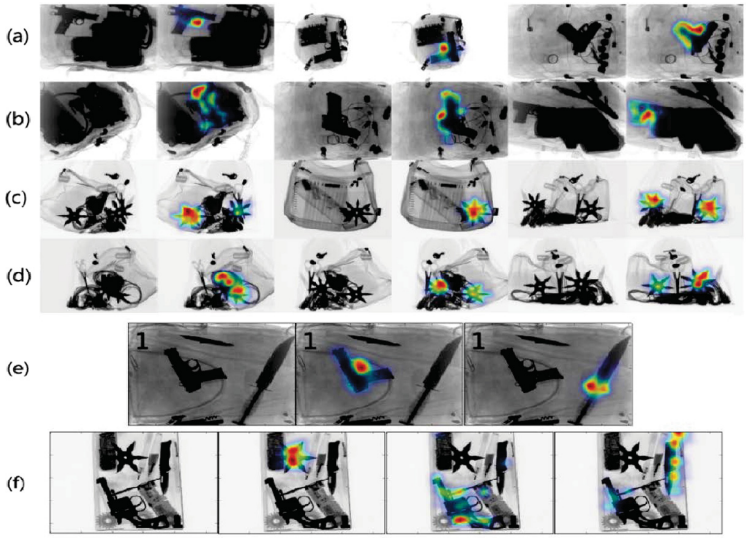


Fig. 7. Results of single-target and multi-target detection. (a), (b) Single-target salience maps. (c), (d) Salience maps of similar multiple targets. (e), (f) Multi-target salience maps.

To quantitatively evaluate the localization effectiveness of our model, we use the above salience map to generate the bounding box, as shown in Fig. 8, which preserves 99% energy of the salience map. Each bounding box in a testing image is compared with the ground-truth bounding box, and the IoU (Intersection over Union) is calculated by Eq. (7). We compare the localization performance of our attention model with the traditional deconvolution method in Table 2.

$$IoU = \frac{area(C) \bigcap area(G)}{area(C) \bigcup area(G)}, \tag{7}$$

where $area(C)$ is the area of candidate bound, and $area(G)$ is the area of ground-truth bound.

Our model can produce highly discriminative salience maps, which is essential for prohibited item detection. Due to the introduction of neuronal stimulus inhibition, the salience maps generated by high-level semantic feedback model

Table 2. Localization IoU of our attention model and the traditional deconvolution method.

Method	Localization IoU
Deconvolution	34.3%
Ours	56.6%

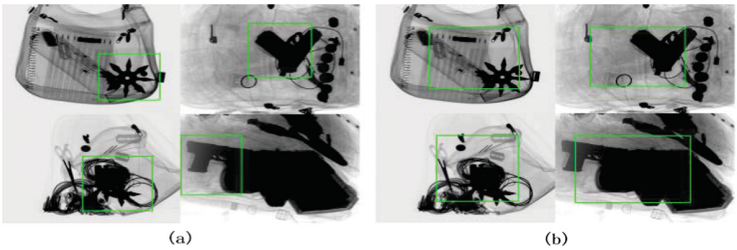


Fig. 8. Results of target localization. (a) The result of our method. (b) The result of deconvolution method.

are highly relevant to the target objects. We get better localization performance than the traditional deconvolution method, as shown in Fig. 8. To evaluate Real-Time performance of our method, we perform the target detection on 2700 X-Ray security images. It takes only 0.76 seconds to process an image on average. So the effectiveness of our method is good enough to meet the needs of real applications.

6 Conclusion

In this paper, we applied an Attention Mechanism based CNN model to achieve detection for prohibited item in airport security X-ray images. It can achieve recognition and location of prohibited item but only need weak supervision training. Our model jointly reasons the outputs of class nodes and the activation of hidden layer neurons during the feedback process. High level semantic is captured and mapped to the image space as an attention map after suppressing noise and interference.

During the inspection process, the CNN can tell the security inspectors the category of prohibited item. At the same time, the attention maps of the prohibited item can remind the security inspectors where the dangerous goods are, facilitating the reinspection. We believe that Attention Mechanism based Convolutional Neural Network provides a new direction for automated security.

Acknowledgments. This work was supported by the National Science Foundation of China Nos. 61379102, 61806208.

References

1. Mery, D., Svec, E., Arias, M., et al.: Modern computer vision techniques for x-ray testing in baggage inspection. IEEE Trans. Syst. Man Cybern. Syst. **47**(4), 682–692 (2017)
2. Mclay, L.A., Lee, A.J., Jacobson, S.H.: Risk-based policies for airport security checkpoint screening. INFORMS (1999)
3. Riffo, V., Mery, D.: Automated detection of threat objects using adapted implicit shape mode. IEEE Trans. Syst. Man Cybern. Syst. **46**(4), 472–482 (2017)
4. Franzel, T., Schmidt, U., Roth, S.: Object detection in multi-view X-ray images. Joint DAGM **7476**, 144–154 (2012)
5. Kundegorski, M.E., Akcay, S., Devereux, M., et al.: On using feature descriptors as visual words for object detection within X-ray baggage security screening. In: International Conference on Imaging for Crime Detection and Prevention 2016, vol. 12, no. 6 (2016)
6. Akcay, S., Kundegorski, M.E., Devereux, M., et al.: Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery. In: IEEE International Conference on Image Processing 2016, pp. 1057–1061 (2016)
7. Cao, C., Huang, Y., Yang, Y., et al.: Feedback convolutional neural network for visual localization and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **99**, 1 (2018)
8. Cao, C., Wang, Z., Wang, L., et al.: Lateral Inhibition-inspired Convolutional Neural Network for Visual Attention and Saliency Detection. Association for the Advancement of Artificial Intelligence (2018)
9. Zhang, J., Bargal, S.A., Lin, Z., et al.: Top-down neural attention by excitation backprop. Int. J. Comput. Vis. **17**, 1–19 (2016)
10. Wang, Q., Zhang, J., Song, S., et al.: Attentional neural network: feature selection using cognitive feedback. In: International Conference on Neural Information Processing Systems. MIT Press, pp. 2033–2041 (2014)
11. Arkachar, P., Wagh, M.D.: Criticality of lateral inhibition for edge enhancement in neural systems. Neurocomputing **70**(4–6), 991–999 (2007)
12. Krell, M.M., Seeland, A., Kim, S.K.: Data augmentation for brain-computer interfaces: analysis on event-related potentials data (2018)