Combining convolutional neural networks and spatial-channel “squeeze and excitation” block for multiple-label image classification

Thuvanan Borvornvitchotikarn, Thongchai Yoyativong
School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand

Article Info

ABSTRACT

In emergency rooms and intensive care units, catheters and tubes are used to keep critically ill patients alive. Appropriate catheter or tube insertion is crucial to avoiding serious complications. Such issues can be rectified if they are identified early. Chest X-rays are commonly used to assess catheter placement. Convolutional neural networks (CNN) have recently been found to enhance multi-label classification tasks on chest X-rays images. Furthermore, attention modules have shown the effect of enhancing spatial encoding on network feature maps. This research analyzed the experiments of each CNN model with different attention blocks. Resnet200D with batch normalization and spatial-channel squeeze and excitation block (BN+scSE) is the best architecture for multiple-label image classification on a chest X-rays dataset from National Institutes of Health Clinical Center (NIH) with multiple catheters and tubes. Then came EfficientNetB5 with BN+scSE and Inception_v3 with spatial squeeze and channel excitation block, respectively.

Keywords:
Attention mechanism
Catheters and tubes
Chest X-rays
Convolutional neural networks
Multi-label image classification

This is an open access article under the CC BY-SA license.

1. INTRODUCTION

Nowadays, catheters or tubes are used to keep critically ill patients alive in emergency departments and intensive care units [1]. Tubes have many advantages. For example, an endotracheal tube (ETT) is used to provide oxygen to the lungs in the case of respiratory issues or lung illnesses. The nasogastric tube (NGT) is useful for individuals who are unable to eat or who consume less than 60% of their daily calorie requirements [2].

Misplacements of tubes can occur in three common tubes. First, in the case of ETT misplacement, the correct ETT position is between 5 and 7 cm above the carina. The Carina position is used for estimating the position of the ETT. The result of ETT misplacement is that accidental intubation of a bronchus is more common on the right main bronchus and the left main bronchus. Second, a chest X-rays is used to determine the position of the NGT. Correct the tip of the NGT position so that it is visible at least 10 cm. beyond the esophagus junction. NGT misplacement that might be passed into the airways. Tube misplacement has many issues, such as looped tube misplacement in the right main bronchus. Finally, a chest X-rays determines the position of the central venous catheter (CVC) position. Correct central venous catheter position is the position of the tip of the CVC within the superior vena cava at or just above the level of the carina. CVC misplacement could lead to many complications, such as pneumothorax or pneumomediastinum [3]. Evidently, the proportions of mispositioned cases encountered have been found to be 5%-28% of ETT, about 15% of NGT, and 2%-7% of CVC, respectively. If catheters are placed incorrectly, severe complications might arise. Commonly, chest X-rays are used to estimate the positioning of various medical equipment and to assess any associated issues.

Journal homepage: http://ijai.iaescore.com
following placement or misplacement [1]. However, each day, radiologists may obtain many patients to radiograph and interpret them, and there may be delays. An assessment of tube position is complicated. Computer-aided techniques have the potential to help radiologists prioritize radiographs with possibly mispositioned catheters for interpretation and to automatically include information suggesting catheter placement in radiology reports, therefore increasing radiologists’ performance [1].

In the literature, the classification of chest X-rays of catheters and tubes has been researched for a long time. The classification methods for tubes have been proposed using various methods, including rule-based [4], and decision trees [5]. Other classification methods based on deep learning (DL) such as, Alexnet [6], GoogLeNet [6], UNET [7], ResNet [8], and EfficientNet [9]. However, these methods are still not optimal in multiple classification tasks, these methods are still inefficient because image classification involves multiple tubes from the current intubation depending upon the patient.

Recent studies on attention mechanisms show that this allows neural networks to change the informative input features, allowing them to maximize network ability and improve model performance. As a result, it has been widely used in neural networks for various machine learning problems in recent years, including, image classification [10]–[12], image segmentation [13], [14], image recognition [15], [16], and natural language processing (NLP) [17], [18]. Attention mechanisms have shown the advantages of enhancing spatial or channel encoding, e.g., squeeze-and-excitation block (SE) [10], bottleneck attention module (BAM) [19], convolutional block attention module (CBAM) [20], spatial squeeze and channel excitation block (cSE) [13], [14], channel squeeze and spatial excitation block (sSE) [13], [14], and spatial and channel squeeze and excitation block (scSE) [13], [14].

In this work, we conduct a study that focuses on the multi-label classification task. We have 11 targets in chest X-rays images, and they can be divided into 4 groups: ETT, NGT, CVC, and Swan. This paper is organized as follows. Section 2 presents the research method. Section 3 describes the experimental results and conclusion will be discussed in section 4, respectively.

2. METHOD

In this section, we will explain this research model for tube image classification, beginning with the implementation and moving on to the attention mechanism and the head layers. Section 2.1 presents the descriptions of the materials in our work. The details about the attention mechanism are presented in section 2.2. The explanation of well-known modules is presented in section 2.3.

2.1. Implementation

All the experiments were trained and tested on Processor Intel (R) Core (TM) i7-2600 CPU @ 3.40 GHz, 3,401 MHz, 4 Core(s), 8 Logical Processor(s), RAM 12.0 GB, with GPU NVidia GeForce RTX3060 12 GB. We utilized the RANCZR-CLiP dataset from the national institutes of health clinical center (NIH), which is available on Kaggle and comprises adult chest X-rays images. The RANCZR-CLiP dataset comprises 30,083 manually labeled images: 24,062 images for the training set, 6,021 images for the validation set, and 3,255 images for the test set [21]. All image data obtained in this experiment are stored in 3×448×448 pixels as shown in Figure 1. The number of each tube type shows in Figure 2. To evaluate our experiments on multiple tubes image classification task, we choose the three types of the most popular convolutional neural networks (CNN) as backbones to conduct this experiment, including Inception_v3 [22], Resnet200D [23], and EfficientNetB5 [24]. Figure 3 shows the placing of attention module in the model of multiple-label image classification.

Figure 1. Examples of chest X-rays images with catheters and tubes [21]
2.2. Attention mechanism

In this section, we will briefly introduce the concept of attention mechanism through the squeeze and excitation block (SE block). SE block was proposed by Hu et al. [10], which was an approach for enhancing the channel relationship. The authors designed the new network to increase a network's representational ability by explicitly modeling the interdependencies between its convolutional features' channels. This block could offer a feature recalibration technique that allows the network to learn to use global information to selectively emphasize informative features and suppress less helpful features. A computational structure of the SE block is presented in Figure 1. For any transformation $T_{tr} = F \rightarrow O$, an input feature $F \in R^{C' \times H' \times W'}$ and an output feature map $O \in R^{C \times H \times W}$. $H$ and $W$ represent the spatial height and width, $C'$ and $C$ are the input and output channels, respectively. Firstly, the input feature $F$ are passed through a squeeze block. Let $K = [k_1, k_2, ..., k_c]$ is set of a 2D spatial kernel and $\times$ represents a convolutional operator. The $T_{tr}$ provides the output feature map as $O = [o_1, o_2, ..., o_c]$, which is calculated using (1).

$$o_c = k_c \ast F = \sum_{n=1}^{n'} k^n \ast f^n, \quad n = 1, ..., n$$

After that, the transformation output $O$ is passed to squeeze global spatial information by using global average pooling. This step can provide the necessary contextual information. Here, the $c-th$ element of vector $S$ is channel-wise statistics, $i, j$ denotes spatial location. It is calculated using (2).

$$s_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} o_c(i,j), \quad c = 1, ..., c$$
Secondary, the squeeze block is followed by an excitation block that tries to completely capture channel-wise dependencies to make use of the information accumulated during the squeeze process. To achieve this goal, these must learn a non-mutually exclusive connection. This second step utilizes the sigmoid activation, which concerns the enable several channels to be highlighted rather than one-hot activation. To reduce the complexity, two fully connected (FC) layers are used to reduce a dimension $D_1$ and adopted ReLU and FC layer $D_2$ to increase the dimension. Where $r$ denotes the reduction ratio (here: $r = 16$), $D_1 \in \mathbb{R}^{7 \times c}$, and $D_2 \in \mathbb{R}^{c \times 6}$. $\delta$ represents the ReLU function, $\sigma$ is Sigmoid function. The final output $\tilde{O}$ of SE block is recalled using the multiply by $F$, which is calculated using (4).

\[
Z = \sigma(D_2 \delta(D_1 S)) \tag{3}
\]

\[
\tilde{O} = z_c \cdot f_c \tag{4}
\]

### 2.3. Head layers

This work will focus on techniques of enhancing the performance of CNN in the context of multiple-label image classification. In the following subsections, we will briefly explain well-known and widely used modules, i.e., SA [25], pre-activation-spatial attention (PAC-SA) [26], cSE [10], sSE [13], batch normalization (BN)+sSE, scSE [13], [14], and BN+ scSE blocks. In our experiment, all these approaches will be used as benchmarks against our method.

#### 2.3.1. Spatial attention block

A spatial-wise attention block ignores the channel information and equally preserves the feature of different channels. $c \times 3 \times 3$ refers to a $3 \times 3$ convolution with the one channel number of the hidden feature map, $c \times 3 \times m$ refers to a $3 \times 3$ convolution with the $m$ channel number of hidden feature map and $m$ is integer number of output channel number needed. BN denotes the batch normalization operation [25]. In our work, we used $m$ those values [64, 32, 8] respectively. The spatial weights map is adaptively generated by this module to properly represent the spatial connection. Spatial attention can be computed as in (5).

\[
\tilde{O}_{SA} = F \times \sigma(\text{BN}(c \times 3 \times 3 \delta(\text{BN}(c \times 3 \times m(F)))))) \tag{5}
\]

#### 2.3.2. Pre-activation-spatial attention block

To reduce the parameter number of network and increase the performance of multiple-label image classification, we introduce to used spatial attention and pre-activation residual network, namely a pre-activation-spatial attention (PAC-SA) block. The PAC-SA block has three BN activations, two ReLU activations, one $3 \times 3$ convolution with eight channel number of the hidden feature map, one $3 \times 3$ convolution with one channel number of the hidden feature map, and one Sigmoid activation. It places BN and ReLU as pre-activation units before convolution operation. The recalibrated values are multiplied by the feature map’s different weights to get the output in (6).

\[
\tilde{O}_{PAC-SA} = F \times \sigma(\text{BN}(c \times 3 \times 3 \delta(\text{BN}(c \times 3 \times 3 \delta(\text{BN}(F)))))\tag{6})
\]

#### 2.3.3. Spatial squeeze and channel excitation block

Roy et al. [14] considered the input feature map $f^n$, cSE block is used to squeeze the spatial information of $f^n$ that is passed by (1) followed by global average pooling in (2). To excite the channel information of the squeezed vector $S$, these can be expressed in (3). The final step shows that the importance of $\tilde{O}_{cSE}$ which is calculated using (7).

\[
\tilde{O}_{cSE} = [\sigma(s_i) f_{i1}, ..., \sigma(s_i) f_{ic}] \tag{7}
\]

#### 2.3.4. Channel squeeze and spatial excitation block

Roy et al. [13] introduced sSE, which compress the information from all channels by convolution $c \times 1 \times 1$. the feature maps of $F$ are squeezed along the channel and are excited the spatial data. Let $F = [f_{i1, j1}, f_{i2, j2}, ..., f_{ij, k}]$, represent the slice on the spatial dimension, where spatial location ($i,j$). In this experiment, we adopted the original sSE as shown in (8) and applied BN after the sigmoid function as shown in (9) for this experiment.

\[
\tilde{O}_{sSE} = F \times \sigma(c \times 1 \times 1(F)) \tag{8}
\]
\[ \hat{\sigma}_{BN+sSE} = F \times \sigma(BN(c^1 \times 1(F))) \]  

(9)

2.3.5. Spatial and channel squeeze and excitation block

Roy et al. [14] presented scSE, which declared cSE see in (7) and sSE see in (8) blocks in parallel and both outputs are added in (10). scSE could generate both channel re-scaling and spatial re-scaling. The feature maps are recalibrated to learn more meaningful feature maps, which are provided both in terms of spatial and channel relevance see in (10). For improving the performance of CNN model and attention block, we proposed BN that normalize the input feature maps and multiplied the output of cSE and sSE blocks. This could get both spatial and channel importance of each location on feature maps cSE and sSE elementwise see in (11).

\[ \hat{\sigma}_{scSE} = \hat{\sigma}_{cSE} + \hat{\sigma}_{sSE} \]  

(10)

\[ \hat{\sigma}_{BN+scSE} = \hat{\sigma}_{cSE} \times \hat{\sigma}_{BN+sSE} \]  

(11)

3. RESULTS AND DISCUSSION

This section will discuss the outcomes of each experiment. Inception_v3 [22], Resnet200D [23], and EfficientNetB5 [24] architecture models were tested on training and validation data. In addition, our experiment should provide a concise and pre-cise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn in comparison with other modules, initial baseline SA [25], PAC-SA [26], cSE [10], sSE [13], BN+sSE, scSE [13], [14], and BN+ scSE blocks. The goal of this experiment is to determine the mean of accuracy score (mAs) and mean of average precision score (mAPs) of each architectural model during the image training process. This experiment will be run using 16 training epochs. Tables 1 to 3 presents the number of network parameters, mAs, and mAPs of each CNN architecture with different attention layers.

As computes the accuracy of the model, the function returns 1, 0 if the subset accuracy is 1, if the entire set of predicted labels matches the true set of labels, otherwise it is 0. In this experiment, we trained CNN models in 5-fold validations \( v \), where \( y'_i \) denotes the predicted value of the \( i \)-th sample, \( y_i \) represents is the corresponding true value, and the fraction of correct predictions overs. The mean of As can evaluate in (13).

\[ A_S = \frac{1}{\sum_{i=0}^{n-1} 1(y'_i = y_i)} \]  

(12)

\[ mAS = \frac{1}{\sum_{a=v}^{\sum_{\alpha}^v} A_{S\alpha}} \]  

(13)

In APs computation, the weighted mean of precision acquired at each threshold is used to summarize a precision-recall curve, with the increase in recall from the previous threshold used as the weight. \( R_n \) and \( P_n \) denote the recall and precision at the \( n \)-th threshold. The mean of APs can calculate in (15).

\[ AP_{S} = \sum_{\alpha}(R_{\alpha} - R_{\alpha-1})P_{\alpha} \]  

(14)

\[ mAP_{S} = \frac{1}{\sum_{\alpha=v}^{\sum_{\alpha}^v} AP_{S\alpha}} \]  

(15)

Table 1 presents the results of Inception_v3 with different attention blocks. cSE block got the best results from all the attention blocks, with mAs 94.68\%, mAPs 65.34\%. Then, followed by the three high performances, including the scSE got mAs 94.70\%, mAPs 65.24\%, the SA got mAs 94.73\%, mAPs 64.96\%, and the BN+sSE got mAs 94.68\%;mAPs 60.01\%, respectively.

<table>
<thead>
<tr>
<th>Inception_v3 with different attention blocks</th>
<th>Model</th>
<th>Parameters</th>
<th>mAs</th>
<th>mAPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception_v3 [22]</td>
<td>baseline</td>
<td>20,227,637</td>
<td>93.76</td>
<td>62.49</td>
</tr>
<tr>
<td>cSE [10]</td>
<td>40,699,371</td>
<td>94.68</td>
<td>65.34</td>
<td></td>
</tr>
<tr>
<td>SA [25]</td>
<td>43,413,939</td>
<td>94.73</td>
<td>64.96</td>
<td></td>
</tr>
<tr>
<td>PAC-SA [26]</td>
<td>43,431,219</td>
<td>94.28</td>
<td>63.14</td>
<td></td>
</tr>
<tr>
<td>sSE [13]</td>
<td>38,601,715</td>
<td>94.66</td>
<td>64.77</td>
<td></td>
</tr>
<tr>
<td>BN+sSE</td>
<td>40,707,563</td>
<td>94.68</td>
<td>66.01</td>
<td></td>
</tr>
<tr>
<td>scSE [13], [14]</td>
<td>40,707,563</td>
<td>94.70</td>
<td>65.24</td>
<td></td>
</tr>
<tr>
<td>BN+ scSE</td>
<td>40,707,571</td>
<td>93.22</td>
<td>61.48</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Number of network parameters, mean of accuracy score, and mean of average precision score of Inception_v3 with different attention blocks.
Table 2 presents the results of the Resnet200D with different attention blocks. BN+ scSE block got the best results from all the attention blocks, with mAs 95.87%, mAPs 72.15%. Then, followed by the three high performances, including the scSE got mAs 95.85%, mAPs 71.31%, the BN+sSE got mAs 95.79%, mAPs 70.46%, and SA got mAs 95.79%, mAPs 70.42%, respectively.

Table 3 presents the results of the EfficientNetB5 with different attention blocks. BN+ scSE block got the best results from all the attention blocks, with mAs 95.24%, mAPs 68.44%. Then, followed by the three high performances, including the scSE got mAs 95.24%, mAPs 68.21%, the BN+sSE got mAs 95.21%, mAPs 68.07%, and PAC-SA got mAs 95.16%, mAPs 68.02%, respectively.

Table 2. Number of network parameters, mean of accuracy score, and mean of average precision score of Resnet200D with different attention blocks

<table>
<thead>
<tr>
<th>Model</th>
<th>Attention module</th>
<th>Parameters</th>
<th>mAs</th>
<th>mAPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resnet200D [23]</td>
<td>baseline</td>
<td>66,862,955</td>
<td>94.31</td>
<td>63.30</td>
</tr>
<tr>
<td>cSE [10]</td>
<td></td>
<td>81,557,867</td>
<td>95.81</td>
<td>69.86</td>
</tr>
<tr>
<td>SA [25]</td>
<td></td>
<td>84,272,435</td>
<td>95.79</td>
<td>70.42</td>
</tr>
<tr>
<td>PAC-SA [26]</td>
<td></td>
<td>84,289,715</td>
<td>95.77</td>
<td>70.32</td>
</tr>
<tr>
<td>sSE [13]</td>
<td></td>
<td>79,460,211</td>
<td>95.72</td>
<td>69.57</td>
</tr>
<tr>
<td>BN+ + sSE</td>
<td></td>
<td>79,460,211</td>
<td>95.79</td>
<td>70.46</td>
</tr>
<tr>
<td>scSE [13], [14]</td>
<td></td>
<td>81,566,059</td>
<td>95.85</td>
<td>71.31</td>
</tr>
<tr>
<td>BN+ scSE</td>
<td></td>
<td>81,566,059</td>
<td>95.87</td>
<td>72.15</td>
</tr>
</tbody>
</table>

Table 3. Number of network parameters, mean of accuracy score, and mean of average precision score of EfficientNetB5 with different attention blocks

<table>
<thead>
<tr>
<th>Model</th>
<th>Attention module</th>
<th>Parameters</th>
<th>mAs</th>
<th>mAPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>EfficientNetB5 [24]</td>
<td>baseline</td>
<td>32,559,675</td>
<td>94.33</td>
<td>63.72</td>
</tr>
<tr>
<td>cSE [10]</td>
<td></td>
<td>47,254,587</td>
<td>95.17</td>
<td>67.09</td>
</tr>
<tr>
<td>SA [25]</td>
<td></td>
<td>49,969,155</td>
<td>94.51</td>
<td>65.29</td>
</tr>
<tr>
<td>PAC-SA [26]</td>
<td></td>
<td>49,986,435</td>
<td>95.16</td>
<td>68.02</td>
</tr>
<tr>
<td>sSE [13]</td>
<td></td>
<td>45,156,931</td>
<td>94.98</td>
<td>66.72</td>
</tr>
<tr>
<td>BN+ + sSE</td>
<td></td>
<td>45,156,931</td>
<td>95.21</td>
<td>68.07</td>
</tr>
<tr>
<td>scSE [13], [14]</td>
<td></td>
<td>47,262,779</td>
<td>95.24</td>
<td>68.21</td>
</tr>
<tr>
<td>BN+ scSE</td>
<td></td>
<td>47,262,787</td>
<td>95.24</td>
<td>68.44</td>
</tr>
</tbody>
</table>

According to these findings, overall, Resnet200D with the addition of the Bn+scSE block outperforms all trained CNN architectural models by mAs 95.87% and mAPs 72.15% see details in Table 2. The EfficientNetB5 architecture followed, with mAs 95.24% and mAPs 68.44% see details in Table 3, and Inception_v3 with mAs 94.68%, mAPs 65.34%. See details in Table 1, respectively.

4. CONCLUSION

This paper has evaluated the experiments of each CNN model with different attention blocks. In order by CNN model, Resnet200D with the improved scSE is the best architecture for multiple-label image classification on a chest X-rays dataset with multiple catheters and tubes from NIH. Then came EfficientNetB5 with improved scSE and Inception_v3 with cSE. This experiment shows the higher performance of the CNN model with the attention block than the baseline CNN model, which could improve the terms of mAs, and mAPs. In addition, we have improved the attention module, i.e., the improved sSE and the improved scSE. They use BN to normalize the input feature maps of these attention blocks, which can reduce covariate shift and reduce the effects of exploding and vanishing gradients.

ACKNOWLEDGEMENTS

This research is part of a PhD dissertation, and the authors would like to thank Mae Fah Luang University for funding the publication of this research.

REFERENCES


Combining convolutional neural networks and spatial-channel ... (Thuvanan Borvornvitchotikarn)
P. Zhao, J. Zhang, W. Fang, and S. Deng, "SCAU weakly


His main research interests are Agricultural Engineering, Data Mining, Artificial Neural Network, and Computer Communications (Networks). He can be contacted at email: thongchai.mfu@gmail.com.