Enhancing Handwritten Digit Recognition using Auxiliary Classifier Generative Adversarial Networks and Self-attention Mechanism

Tingkai Hu

Stony Brook University, Anhui University, Anhui, Hefei, 230039, China

Abstract. This research study investigates the integration of the self-attention mechanism and the Auxiliary Classifier Generative Adversarial Networks (ACGAN)to improve handwritten digital recognition using the MNIST data set. Although progression has been made in the generative Adversarial Networks (GANs) used for image synthesis, it is still tested to attain top-notch, context-accurate photo generation, specifically under various information problems. This study fixes these spaces by incorporating self-attention with ACGANs, improving the integrity and labelling accuracy of the produced images. This approach entails changing the ACGAN framework to incorporate the self-focus module, so far better identify the context dependencies in the photo. This assimilation advertises much more detailed and accurate numerical depiction, which is especially helpful when converting data sets with dark histories right into more clear, segmented white background pictures, which enhances the differentiation between data classes. The outcomes reveal that the recognition accuracy and processing speed have been significantly boosted, validating the adaptability and usefulness of the design in various procedure circumstances. The research results show that this method can establish brand-new standards for producing high-grade electronic photos in numerous applications and show considerable progress in the field of image recognition technology.

1. Introduction

The fast development of deep learning technology has significantly boosted the capacity of the image recognition system. Generative adversarial Networks (GAN) have ended up being a vital innovation to generate real data and advertise complex image recognition jobs. Nevertheless, typical GAN designs commonly experience stability troubles, such as pattern collisions, specifically when managing restricted or out-of-balance data collections (such as Modified National Institute of Standards and Technology database (MNIST) data sets).

Limitations of typical GAN approaches: Traditional GAN frequently has problems with pattern collapse and fails to catch the intrinsic diversity of data collections such as MNIST. [1]. This normally results in a non-diversified outcome, which limits the effectiveness of the design in applications that require high irregularity and precision in image recognition.

Corresponding author: TingkaiHu@cqu-edu.cn

Presenting new techniques: To meet these obstacles, current innovations have presented improved versions, such as Self-Attention Generative Adversarial Network (SAGAN) and Auxiliary Classifier Generative Adversarial Networks (ACGAN) [2, 3] SAGAN combines the self-attention mechanism, which assists in capturing remote dependencies in images and significantly improves the uniformity and information of the created pictures. On the other hand, ACGAN uses complementary classifiers to impose label consistency throughout the training procedure, so regarding make sure the accuracy and diversity of the generated photos.

Empirical efficiency improvement: Empirical research study reveals that SAGAN not only boosts the Inception Score but likewise efficiently lowers the Frechet Inception distance on complicated data sets such as ImageNet, which shows that it can be observed on MNIST. Similar advantages. This performance enhancement shows the improvement of image quality and design security, which is essential for efficient training and classification precision.

Improvement recap: The combination of the self-conference device and auxiliary category within the GAN structure notes substantial development in conquering the limitations of the early GAN model. These improved GAN designs have been verified to produce more trusted and diverse photo results, considerably improving the robustness needed for sensible applications, including the digital category on MNIST data collections.

Research study emphasis: This study concentrates on how to integrate the self-attention mechanism with ACGAN to enhance the generation of diverse and high-grade images and enhance the performance of handwritten digital recognition models utilizing MNIST information sets. By optimising the generation process, this approach aims to dramatically enhance the toughness and precision of the classification design trained in this benchmark information established.

2. Dataset

2.1 MNIST dataset

The MNIST data set comes from the modified database of the National Institute of Standards and Technology, which is an essential benchmark for machine learning and computer system vision study. The data set consists of 70,000 greyscale pictures of transcribed numbers from 0 to 9. Each picture is formatted as a 28x28 pixel grid, in which the worth of each pixel ranges from 0 (black) to 255 (white), indicating various levels of greyscale intensity.

The MNIST data set was originally developed by a blend of NIST's special database 1 and unique database 3, including transcribed numbers from American high school students and U.S. Census Bureau staff members. These beginnings are essential because they guarantee the diverse depiction of handwriting designs, which is necessary for the growth of effective devices finding out versions, which are good at handling real-world adjustments in handwriting digital acknowledgment.

2.2 Structure of the MNIST dataset

Educating set: It consists of 60,000 photos and is used to educate artificial intelligence versions to precisely recognize and categorize transcribed numbers.

Test set: including 10,000 images to assess the efficiency and accuracy of the training version in a consistent and standard means.

The prevalent use of MNIST information embedded in the academic community and research study is due to its duty as a reliable training and testing room for the advancement and benchmarking of machine learning algorithms. Its standardisation allows researchers to straight compare the effectiveness of brand-new computer modern technologies and offers

standard information established for checking out advanced machine learning and picture processing techniques [4].

3. Research method this paper chooses

3.1 self-attention Gan

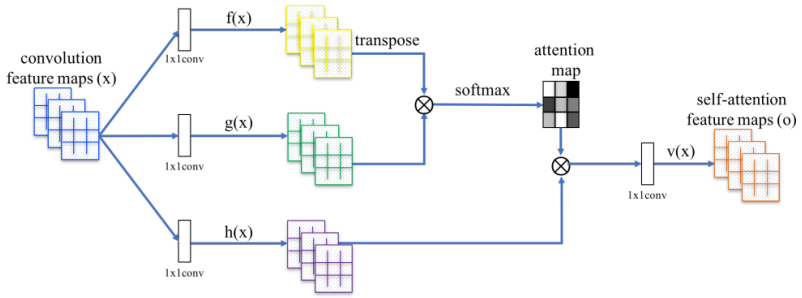


Fig. 1 The flow chart of self-attention Gan [5].

3.1.1 Self-attention module

Fig. 1 shows the self-attention components utilized in structures such as the SAGAN [6]. This procedure can be summarised as follows:

Input Feature Maps (x): This component takes a collection of attribute mappings from the previous convolutional layer as input.

Transformations: The input feature mapping xxx has undertaken 3 independent 1x1 convolutional transformations:

$f(x)$: Produces a transformed feature map, which is then transposed.

$g(x)$: Generates another transformed map, retaining spatial information.

$h(x)$: Creates a third set of transformed feature maps.

Attention Map: The outputs $f(x)$ and $g(x)$ are matrix-multiplied, followed by a softmax operation to generate an attention map. This map highlights the value of various locations in the image relative to each other [7].

Self-Attention Feature Maps: The focus map is utilized to weight the feature map, hence generating a self-contention feature map, which integrates the context details of the entire photo [8].

Output: The final result is a set of fine-tuned function diagrams, which is given by the global context of the photo, which boosts the ability of the design to produce pictures with much better comprehensibility and information [8].

This self-attention mechanism permits the model to concentrate on the appropriate areas of the photo, therefore boosting the overall image quality and the consistency of the produced outcome.

3.1.2 Self-attention layer

The self-attention layer allows the model to focus on different parts of the input information by capturing the partnership in between remote features [7]. Unlike the typical layer that only considers regional details, self-attention allows the design to weigh the significance of each

part of input about various other parts. This worldwide perspective helps to create more systematic and contextually accurate pictures because the model can focus on the pertinent locations of the image at the same time. In image generation, this mechanism is particularly helpful for improving the top quality and consistency of created photos [8].

3.2 ACGan

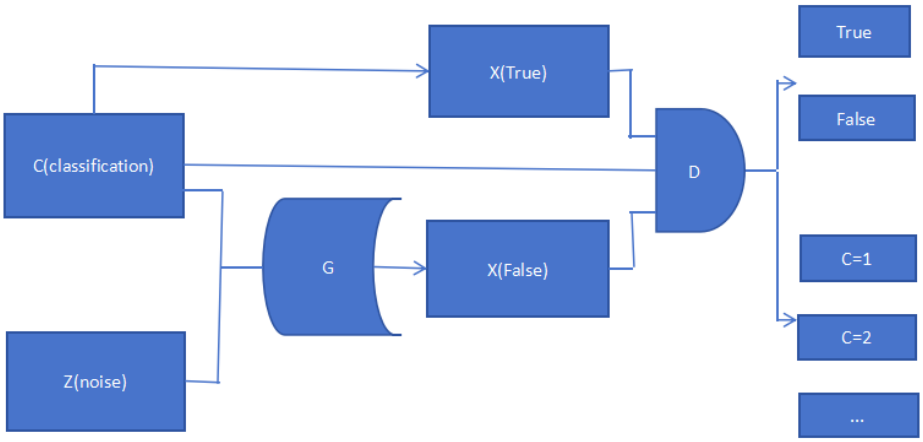


Fig. 2 The flow chart of the auxiliary classifier Gan (Photo/Picture credit: Original).

Fig. 2 illustrates the architecture of an ACGAN [6]. In this model, the Generator (G) takes a noise vector Z and a class label C as inputs to produce a synthetic image X_{gen}. The Discriminator (D) has a dual role: it distinguishes between real images X_{real} and generated images X_{gen}, and forecasts the class label C of the genuine picture and the created image. This double outcome allows ACGAN to create images that are both genuine and properly labelled, which improves the control and interpretability of the produced result.

3.2.1 Generator

The generator is in charge of creating synthetic information examples similar to genuine information [9]. It takes a prospective vector (generally tasted from the normal distribution) as the input and converts with a collection of layers to produce outcomes that mimic real data attributes. In our application, the generator includes several completely linked layers, complied with by activation functions such as LeakyReLU and Tanh to make certain the nonlinearity and correct scaling of the output.

3.2.2 Discriminator

The discister is a binary classifier that can compare real and created (phony) data examples. It takes the picture as the input, processes it with a series of layers, and outputs a chance to suggest whether the input holds true or false. Additionally, our discister consists of an auxiliary classifier that can anticipate the labels of input information and make it applicable to ACGAN [10].

3.2.3 Loss function

Generator Loss:

The generator loss in ACGAN consists of 2 terms that show its dual goals: producing realistic photos and appropriately classifying these pictures:

$$Generator\ Loss = -\mathrm{E}[\log(D_{\text{real/fake}}(G(z,y)))] - \mathrm{E}[\log(D_{\text{class}}(G(z,y)))] \quad (1)$$

$G(z, y)$: This suggests the output of the generator. Z is the input noise vector, and y is the label vector used to manage the generation procedure. G accepts these inputs and generates a photo that tries to mimic the real image corresponding to the label y.

$D_{\text{real/fake}}(G(z,y))$: This is the result possibility of the distinguisher, that is, the picture produced by G is real (not fake). This term motivates generators to produce practical photos.

$D_{\text{class}}(G(z,y))$: This is the result probability of the disermentator, that is, the photo produced by G comes from the proper class based upon y. The term makes sure that the photos created by the generator are not just practical, but also properly identified according to their tags.

Discriminator Loss:

The discisterer in ACGAN additionally has a double goal: to appropriately identify real and incorrect photos and correctly categorize actual photos:

$$\begin{aligned} Discriminator\ Loss = &-\mathrm{E}[1-\log(D_{\text{real/fake}}(G(z,y)))] - \mathrm{E}[\log(D_{\text{class}}(G(z,y)))] \\ &- \mathrm{E}[\log(D_{\text{class}}(x_{\text{real}}))] - \mathrm{E}[\log(D_{\text{real/fake}}(x_{\text{real}}))] \end{aligned} \quad (2)$$

x_{real} : These are the actual pictures in the information set.

$D_{\text{real/fake}}(x_{\text{real}})$: This term indicates the possibility that the real data established photo of the distinguisher is real. This motivates the distinguisher to correctly determine the actual image. It plays a crucial duty in educating the distinguisher to identify actual information from the created data and enhances its capability to generalise real-world information.

$\mathbf{G}(\mathbf{z}, \mathbf{y})$: This represents the image generated by the generator. Z is the input noise vector, and y is the label vector that readjusts the generation process.

$1 - D_{\text{real/fake}}(G(z,y))$: This term computes the possibility that the identifier correctly identifies the picture created by G as a phony photo. This motivates the distinguisher to effectively distinguish between actual photos and generate images, which is essential for training powerful generation designs.

$D_{\text{class}}(x_{\text{real}})$: This reflects the discriminator's probability that the real image x_{real} is correctly classified according to its true label. The exact classification of genuine pictures ensures that the discister is not only effective in detecting genuine and fake images, but additionally in correctly classifying real photos. This dual function is vital for applications that call for accurate labelling precision.

$D_{\text{class}}(G(z,y)))$: The term examines the capability of the discister to inaccurately categorize the generated photos to make certain that it can recognise and punish the produced images significantly erroneously. It is extremely important to show the discriminant not just to locate fakes, but also to improperly criticise the details characteristics or courses assigned to the created image.

4. Result

4.1 Loss curve analysis

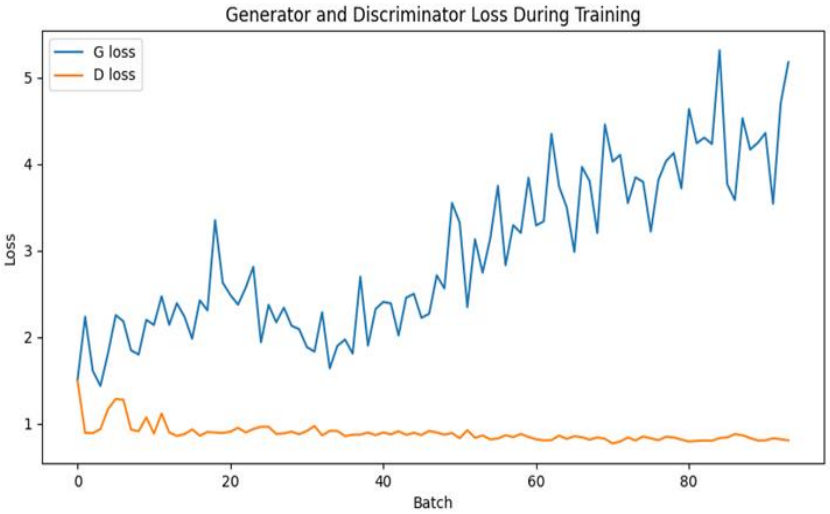


Fig. 3 Loss function of discriminator and generator (Photo/Picture credit: Original)

During the training, the loss of generators (as displayed in Fig. 3) slowly raised, which is very closely related to the integration of the self-attention mechanism into the GAN structure. The self-attention mechanism boosts the capacity of the version to capture remote dependencies in between different locations of the picture, enabling it to handle extra complex image generation tasks and enhance the overall image quality.

Rapid Convergence in Early Training:

At the beginning of training, the interaction between the generator and the distinguisher is stronger, permitting the generator to swiftly generate pictures that match particular designs and tags to complete the first stage of picture training. At this time, the loss of the generator is still relatively reduced, which suggests that it successfully utilizes the feedback from the disaster to create pictures that conform to the target style [3].

Increase in Generator Loss:

With the advancement of training, the loss of the generator increases significantly, which foreshadows the activation of the self-attention mechanism. Right now, the design starts to concentrate on more facility information in the picture, such as the adjustment of background colour and the delimitation of boundaries. The self-attention mechanism particularly enhances the lower area of the image, which can adjust the brightness and contrast, make the history transition from dark white to bright white, and make the boundaries sharp and clear. This shows that via more extensive training, generators are not just efficient in making realistic pictures, but additionally proficient at refining picture information.

Research reveals that the self-attention mechanism significantly boosts the clearness and details of the image and improves its total aesthetic charm. In addition, with the development of the training, the resemblance between the generated pictures and the actual pictures increases, specifically in the border meaning and various other elements. The model shows greater precision, accomplishing stronger contrast and finer information.

These developments aid generators produce images that show the real picture extra closely. In the innovative stage of training, the created pictures show almost the same level of detail and framework as the initial information set.

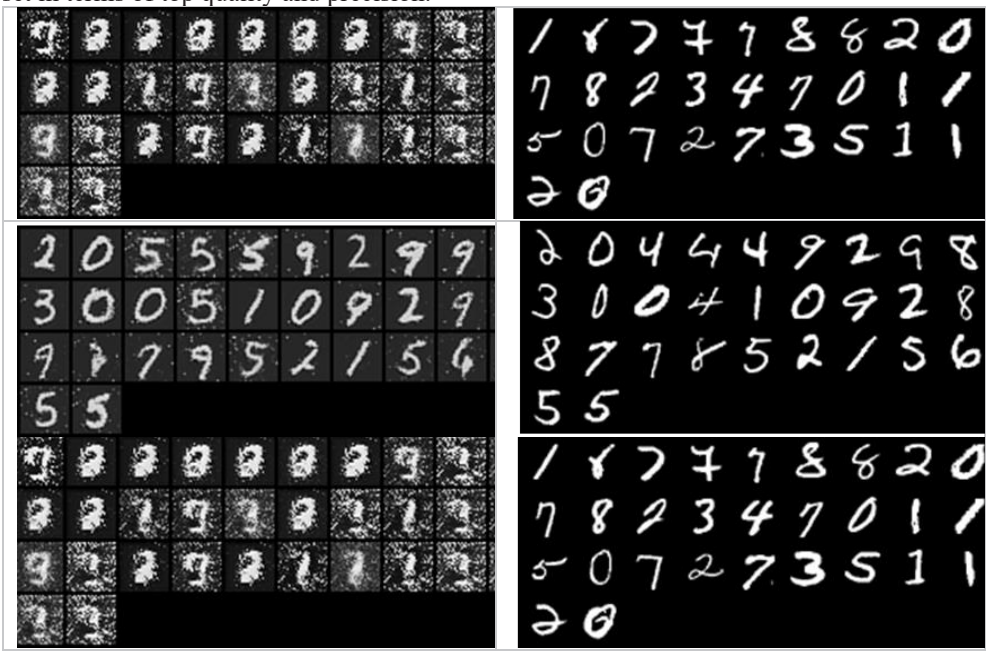
4.2 Generated vs. real image comparison

In Fig. 4, the generated image (left) is juxtapositioned with its real corresponding photo (right) at different training stages. With the development of training, the clearness of the generated photo and the enhancement of the interpretation are evident. In the beginning, these pictures might look blurry and lack thorough features, but as the training continues, they slowly become clearer and extra exact.

This progression is mainly because of the generator's capability to enhance its result via feedback obtained from the distinguisher. The integrated self-attention mechanism plays an essential duty in this enhancement, allowing the generator to focus on the vital areas of the image, thus improving the total image quality. As an example, the sides of numbers ended up being clearer, and the contrast between numbers and their histories is boosted, making the generated pictures more very closely imitate the actual photo.

Through deeper training, the background of the generated photo will certainly also transform from a darker darkness to a brighter white. This change may be influenced by the self-attention mechanism, which triggers generators to concentrate a growing number of on boosting brightness and contrast. This modification to a white background enhances the exposure of numbers, specifically in dark conditions, revealing the useful benefits of the design.

Constant improvement throughout the training procedure enables the generator to create photos that are aesthetically constant with the real instance and preserve consistency in framework and information. By the final stage of the training, the generated pictures have attained a high level of clearness and precision, and closely matched the initial information set in terms of top quality and precision.



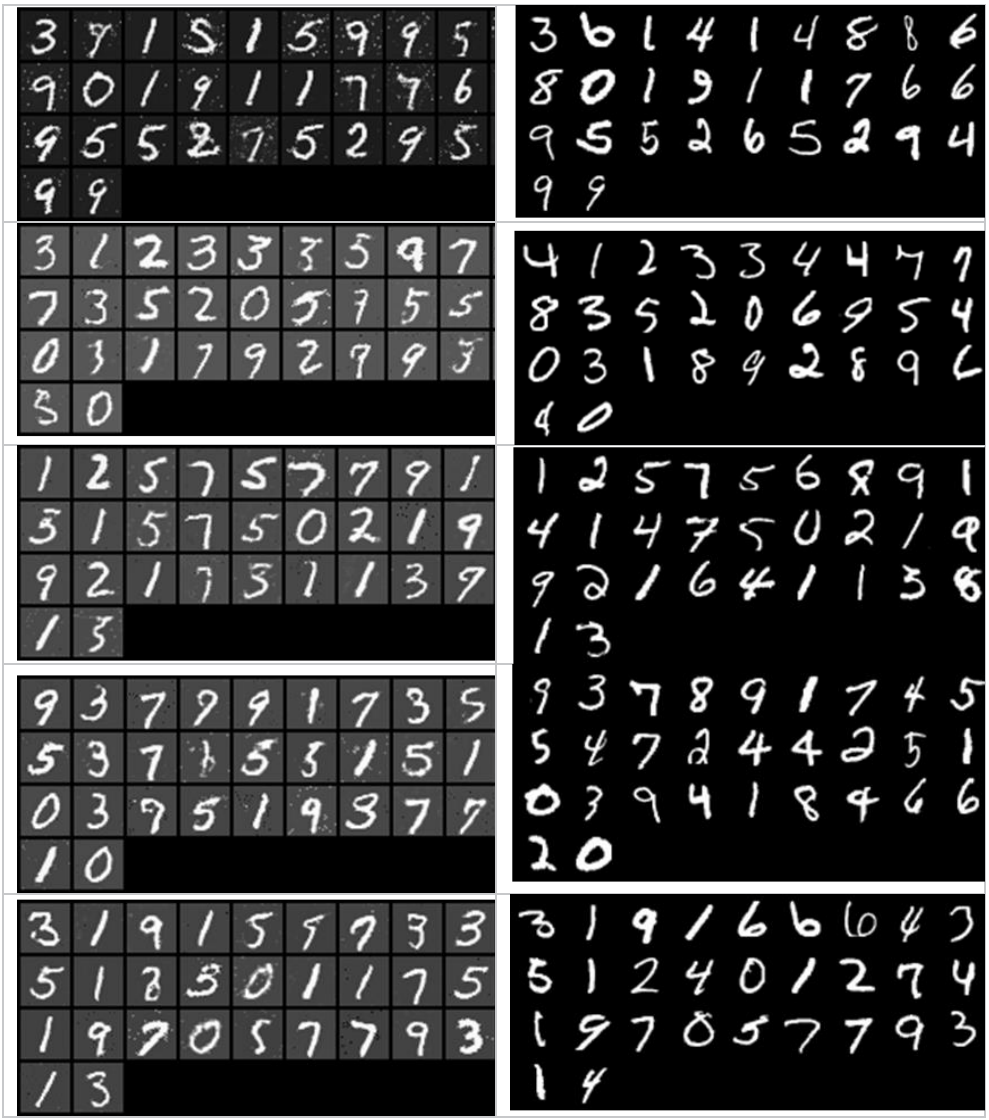


Fig. 4 Progress of the training (Photo/Picture credit: Original).

5. Conclusion

This research concentrates on integrating the self-attention mechanism with the ACGAN to improve the efficiency of handwritten digital recognition. ACGAN is famous for its conditional image generation function. Its generator intends to create very reasonable images that very closely match the defined label, and its distinguisher assesses the authenticity of the photo and the precision of its label at the same time.

The integration of the self-conference mechanism allows the version to better handle long-distance reliances in photos and focus on far-off however appropriate components. This enhancement is essential for tasks that call for premium photo generation and accurate label alignment, especially when slight distinctions might significantly impact the overall accuracy.

With the increase of self-continuation, the design needs fewer training iterations to attain the best effect, and the resolution of the generated photos is substantially enhanced. As the training progresses, the boundaries between numbers become clearer, making images a lot more detailed and easier to identify - which is the key to enhancing recognition accuracy, especially in environments with complicated backgrounds or low contrast.

Furthermore, the version's capability to readjust the photo background from dark to intense includes useful value, which is specifically beneficial in low-light or night conditions, because the contrast needs to be boosted for far better acknowledgment. This capability to dynamically adjust the history verifies the broad applicability of this integrated method in real-world scenarios.

As a whole, the combination of ACGAN and the self-attention mechanism notes substantial progress in conditional image generation technology. This extensive approach not only improves the precision and efficiency of picture generation but also reduces the training time and enhances the adaptability to different image recognition situations. The potential of this technique surpasses handwriting digital recognition and encompasses a bigger variety of applications such as image classification and object detection, paving a brand-new means for top-notch image generation and accurate category efficiency.

References

1. Y. Zhang, et al. Self-attention generative adversarial networks. Proceedings of Machine Learning Research, 119, 7354-7363 (2020).
2. I. Goodfellow, et al. Generative adversarial nets. Journal of Machine Learning Research, 3, 2672-2681 (2021).
3. A. Odena, et al. Conditional image synthesis with auxiliary classifier GANs. Proceedings of the 34th International Conference on Machine Learning, 70, 2642-2651 (2021).
4. Y. LeCun, C. Cortes, C. J. C. Burges. The MNIST database of handwritten digits.
5. H. Zhang, I. Goodfellow, D. Metaxas, A. Odena. Self-attention generative adversarial networks. In International Conference on Machine Learning, 7354-7363. PMLR (2019).
6. A. Odena, C. Olah, J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In International Conference on Machine Learning. PMLR, 2642-2651 (2017).
7. Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi. "On the relationship between self-attention and convolutional layers." *arXiv preprint arXiv:1911.03584* (2019).
8. A. Vaswani. Attention is all you need. arXiv preprint arXiv:1706.03762 (2017).
9. A. Dash, J. C. B. Gamboa, S. Ahmed, et al. Tac-GAN: Text conditioned auxiliary classifier generative adversarial network. arXiv preprint arXiv:1703.06412 (2017).
10. X. Xia, R. Togneri, F. Sohel, et al. Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection. IEEE Transactions on Multimedia, 21(6), 1359-1371 (2018).