Científica, vol. 26, núm. 2, pp. 01-09, julio-diciembre 2022, ISSN 2594-2921, Instituto Politécnico Nacional MÉXICO
DOI: https://doi.org/10.46842/ipn.cien.v26n2a02

Recognition of Facial Expressions Using Vision Transformer

Reconocimiento de expresiones faciales con vision transformer

Paula Ivone **Rodríguez-Azar**[1]
José Manuel **Mejía-Muñoz**[2]
Carlos Alberto **Ochoa-Zezzatti**[3]

Universidad Autónoma de Ciudad Juárez, MÉXICO
[1]https://orcid.org/0000-0001-8981-5350 | al206578@alumnos.uacj.mx
[2]https://orcid.org/0000-0002-5832-6623 | jose.mejia@uacj.mx
[3]https://orcid.org/0000-0002-9183-6086 | alberto.ochoa@uacj.mx

Recibido 03-05-2022, aceptado 19-09-2022

Abstract
The identification of emotions through the reading of non-verbal signals, such as gestures and facial expressions, has generated a new application in the field of Facial Expression Recognition (FER) and human-computer interaction. Through the recognition of facial expressions, it would be possible to improve industrial equipment by making it safer through social intelligence that has excellent applications in the area of industrial security. That is why this research proposes to classify a series of images from the database called FER-2013, which contains data on seven different emotions, which are anger, disgust, fear, joy, sadness, surprise, neutral. For the recognition of expressions, a Vision Transformer architecture was implemented, of which 87% precision was obtained, while the top test accuracy was 99%.

Index terms: vision transformer, facial expressions, emotion recognition.

Resumen
La identificación de emociones a través de la lectura de señales no verbales, como gestos y expresiones faciales, ha generado una nueva aplicación en el campo del Reconocimiento de Expresión Facial (FER por sus siglas en ingles) y la interacción humano ordenador. A través del reconocimiento de expresiones faciales, sería posible mejorar los equipos industriales haciéndolos más seguros a través de la inteligencia social que tiene excelentes aplicaciones en el área de la seguridad industrial. Es por ello que en esta investigación se propone clasificar una serie de imágenes de la base de datos denominada FER-2013, que contiene datos sobre siete emociones distintas, las cuales son enfado, asco, miedo, alegría, tristeza, sorpresa, neutral. Para el reconocimiento de expresiones se implementó la arquitectura Vision Transformer, de la cual se obtuvo un 87% de exactitud, mientras que la exactitud más alta fue de 99%.

Palabras clave: vision transformer, expresiones faciales, reconocimiento de emociones.

I. INTRODUCTION

Emotions are inherent to the human being; they are generated by reacting to various external stimuli. Facial expressions are generally instantaneous subconscious manifestation that represents an emotion. It should be noted that human beings communicate 55% through facial expressions, 38% by intonation, and only 7% by spoken words [1]. FER is useful in the detection of emotions in the industry, this application is useful to prevent accidents in high-risk machinery. About 868,000 work accidents occur daily [2]. One of the main causes is the emotional state of the employees, which causes errors when executing the assigned task [3]. Techniques to detect emotions for accident prevention have been widely used, in [4] a correlation was made between facial expression and occupational accidents. In addition, other studies have analyzed the risks of accidents related to emotions, such as the case of [5], [6].

FER work through the visual information provided by the deformation of facial features, to classify them into various categories [7]. Among the steps to follow for FER, is the preprocessing of the image, which consists of most of the resizing and normalization of the intensity, pixels. In addition to the increase and the classification of the data [8]. Recently, research has used machine learning algorithms to extract and classify different emotions, among which convolutional networks (CNN) stand out.

In FER some studies have obtained favorable results, such as the case of [1] which used the Haar Wavelet Transform (HWT) for the extraction of characteristics and the classification (SVM). For his part [9] used the WiSARD network, also [10] used Neighborhood Difference Features for the extraction of characteristics and random forest for the classification. For the analysis and classification of images, recent studies have focused on the implementation of the vision transformer model [11], [12], [13], [14], [15], [16]. Also, various studies have used the FER-2013 database for testing, for example in [17] and [18] they used a 2D convolutional network and obtained an accuracy of 66% and 94% respectively. In addition, [19] and [20] used the CNN-XCEPTION network and obtained an accuracy of 71% and 65.97% respectively. For their part [21] with Support Vector Machine (SVM) obtained 63% accuracy. On the other hand, in [22] and [23] convolutional networks obtained an accuracy of 69% and 66% respectively. The challenges of FER consist mainly in reading the expressions in the different positions [11]. However, some databases contain enough information for testing and implementing algorithms before taking them to the real world.

In this research, to reduce work accidents, we propose to use the database called FER-2013, which contains photographs of human faces in gray scales, which represent the emotions of anger, disgust, fear, happiness, sadness, surprise, and neutral. The model called vision transformers, which is an adaptation of the previously proposed transformers is used for the classification of the images in the database.

The rest of the paper is organized as follows; in section 2 the proposed methodology will be analyzed. Section 3 shows the results obtained by the model. Section 4 discusses the results obtained by the model. Finally, section 5 shows the conclusions of the proposed model.

II. DEVELOPMENT

A. Proposed model

The model used in this research is the one proposed by [24] called vision transformer (ViT), which is based on the original model Transformer [25]. The ViT consists of dividing each image into segments of the same dimension called patches, arranging them linearly, adding positional inlays, and subsequently entering the information into a conventional transformer. The transformer classifies the characteristics of the images in the different classes. Figure 1 shows the graphic representation of the ViT model. The figure shows the

segmentation of the image in patches, which become vectors pass through the transformer which is made up of a Multi-Layer Perceptron (MLP) network, the attention, and the embedded patches are created.

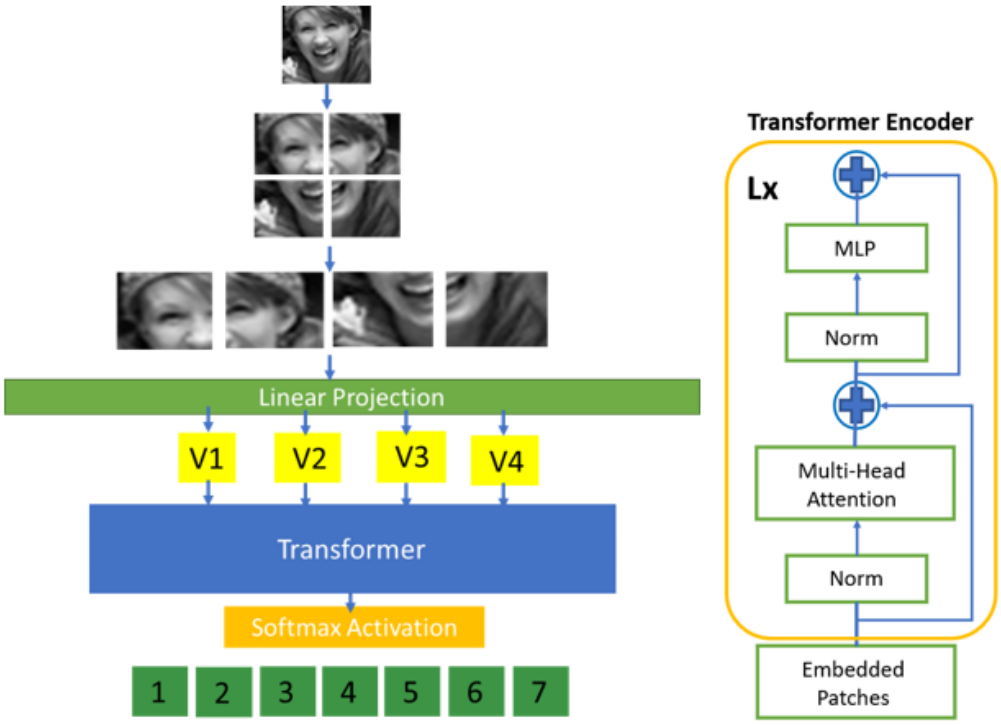

Fig. 1. ViT model sequence. The illustration is inspired by ViT [20].

A conventional transformer consists of multi-head self-care layers, it also goes through Multilayer Perceptron (MLP), and Layer norm (LN) for each block. The ViT reshapes the image $x \in R^{H\times W\times C}$ into a sequence of flattened 2D patches $x_p \in R^{N\times(P^2\cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. Using equation 1 the patches are flattened and mapped to D dimensions with a linear projection. A learnable embed is prepended to the sequence of embedded patches $(z_0^0 = x_{class})$. The equation 2 represent the Multihead Self-Attention (MSA) and the equation 3 a (MLP). The image is representing by equation 4, (z_L^0) is the state at the output of the transformer encoder.

$$Z_0 = \left[x_{class}; x_\rho^1 E; \ldots; x_p^N E\right] + E\rho os,\ \ E \in \mathbb{R}_9^{(p^2\cdot C)\times D},\ E_\rho \in \mathbb{R}^{(N+J)\times D} \tag{1}$$

$$Z_l' = MSA\big(LN(zl-1)\big) + z_{l-1,} \qquad l = 1 \ldots L \tag{2}$$

$$Z_l' = MSA\big(LN(zl-1)\big) + z_{l-1,} \qquad l = 1 \ldots L \tag{3}$$

$$y = LN(Z_L^0) \tag{4}$$

B. Experimental protocol

The FER-2013 data set was used for the experiment (from https://www.kaggle.com/), which has a total of 35,887 grayscale images. The size of each image is 48x48 pixels, 56,630 images were used for training and 6,783 images were used for validation. Fig. 2 shows the examples of the input images by class, a characteristic of this database is that it has images that show expressions in different postures, front, side, different angles of vertical and horizontal rotation. This allows you to get closer to the postures that are presented in a real application.

Fig. 2. FER-2013 database by class.

For this experiment, the data set was randomly divided into 90% for training and 10% for validation. During preprocessing, images are resized and converted to green scale. Figure 3 shows the printout of one of the input images.

Fig. 3. Input Images from FER-2013.

For preprocessing, the images were normalized, resized to 72 x 72. Finally, the dataset was augmented with a horizontal flip and with a rotation interval of [-0.02, 0.02]. For the transformer architecture, Table 1 shows the hyperparameters used for the model. The hyperparameters are based on the adjustments made by [24], where the learning_rate and way_decay are based on the parameters suggested for the optimized Adam. The dimension of the image and the quantity of the patches were obtained experimentally.

TABLE 1
HYPERPARAMETERS FOR THE TRAINING

	PARAMETER	VALUE
1	LEARNING_RATE	0.001
2	WEIGHT_DECAY	0.0001
3	BATCH_SIZE	256
4	NUM_EPOCHS	90
5	IMAGE_SIZE	72
6	PATCH_SIZE	6
7	PROJECTION_DIM	64
8	NUM_HEADS	4
9	TRANSFORMER_LAYERS	2
10	MLP_HEAD_UNITS	[2048, 1024]

The model was fed with patches of the input image. Each patch measured 72x72 pixels, each patch was 6x6 pixels, so 144 patches were obtained per image. Fig. 4 shows one of the images and the creation of its patches.

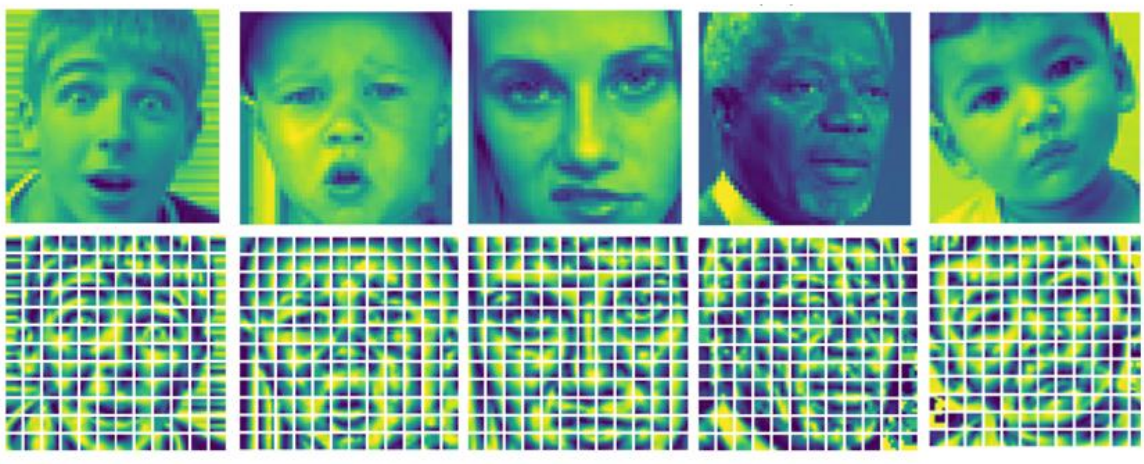

Fig. 4. Patches per image.

Subsequently, the patches were normalized using a norm layer. Finally, they enter the main classifier to be trained and evaluated.

C. Results

For the evaluation of the experiment, the training accuracy and validation accuracy metrics were used. Fig. 5, shows 14 epochs of the training phase, the accuracy curves of the graph show that both magnitudes were increased until reaching the Test accuracy of 87.29%.

In addition, the training loss graph and the validation loss graph were used, this shows how the losses were decreasing which can be seen in Fig. 6.

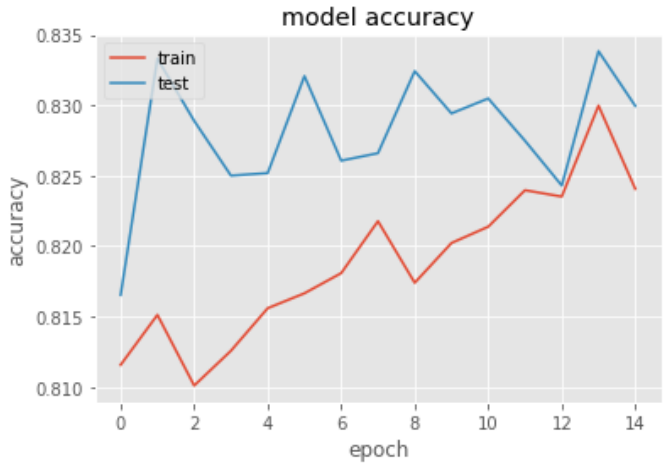

Fig. 5. Model accuracy.

Fig. 6. Model loss.

Another metric that was used in the evaluation was the calculation of the multiclass confusion matrix, which shows us the classifications of the seven emotions to be classified, which are anger, disgust, fear, happy, neutral, sad, and surprise. In the confusion matrix, we can calculate the precision-recall and the total accuracy. For the calculation of recall and precision, we use equations 5, 6 and 7.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FP} \quad (7)$$

Where TP is the true positive, TN is the true positive, FP is the false positive, and FN is the false negative of all classes in the multiclass confusion matrix. Fig. 7 shows the confusion matrix of the FER-2013 database classification using the ViT model.

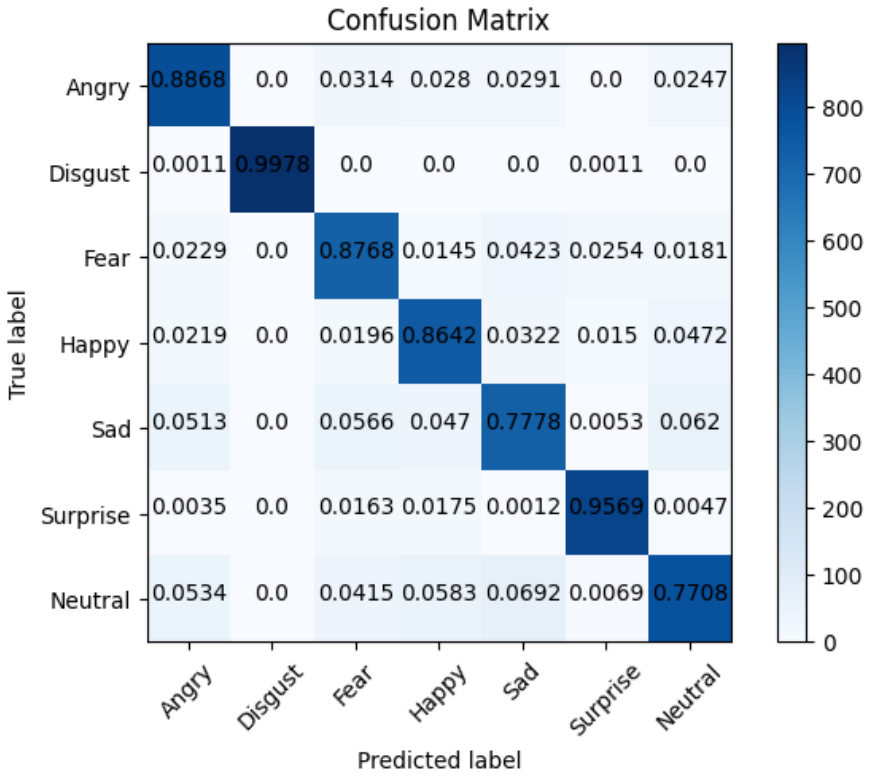

Fig. 7. Confusion matrix from the FER-2013 database.

The crossed results of the evaluation of the algorithms with the FER-2013 database are shown in Table 2, in which it is observed that the ViT algorithm has been superior to those used with convolutional networks.

TABLE 2
CROSS-DATASET EVALUATION RESULTS ON FER-2013.

	METHOD	ACCURACY
1	2D-CNN	66%
2	CNN-XCEPTION	71%
3	CNN-XCEPTION	65%
4	SVM	63%
5	CNN	69%
6	CNN	66%
7	VIT (OURS)	87%

III. CONCLUSIONS

In this research, a classic model called Vision Transformer was presented, which was applied to the database called FER-2013. The database has facial recognition images of seven different emotions which are anger, disgust, fear, happiness, sadness, surprise, and neutral. The overall accuracy of the proposed model was 87%.

In Figure 5, validation accuracy increased throughout training in proportion to training accuracy. While the loss graph shows that the validation losses decreased along with the training losses until they reached less than 0.4.

Regarding the confusion matrix, it is observed that the accuracy obtained by the angry class was 83%, 99.7% disgusted, 86% afraid, 86% happy, 80% sad, 94% surprised, 76% neutral. Indicating that the maximum precision is in disgust, with 99%, that is, less than 1% of the data was confused with anger. The second-highest ranked class was a surprise at 94%, however, this class was confused with all the others. Finally, the classes with the least precision were sadness and neutral, as visually observed, the faces are similar between these last classes.

The accuracy was higher in some classes than in others, with a 24% difference between them. In addition, it is observed that FER in the FER-2013 database using ViT considerably improved the precision over the methods with convolutional networks. However, it is proposed for future work to carry out a real-time validation of the model. As well as to change the hyperparameters of the ViT algorithm, mainly modifying the number of patches and the number of neurons, and finally use other databases to verify the results obtained.

REFERENCES

[1] C. V. R. Reddy, K. V. K. Kishore, "Facial emotion recognition using NLPCA and SVM", *Traitement du Signal*, 2019, vol. 36, no 1, pp. 13-22.

[2] International Labour Organization, https://www.ilo.org/global/lang--en/index.htm

[3] R. N., Sari, D. S. Dewi, "Development models of personality, social cognitive, and safety culture to work accidents in the chemical company", *In IOP Conference Series: Materials Science and Engineering IOP Publishing*, 2021 vol. 1096 no. 1, pp. 012-025.

[4] Y. Kong, H. F. Posada-Quintero, M. S. Daley, K. H. Chon, J. Bolkhovsky, "Facial features and head movements obtained with a webcam correlate with performance deterioration during prolonged wakefulness", *Attention, Perception, & Psychophysics*, 2021, vol. 83, no. 1, pp. 525-540.

[5] A. S. Kumar, N. Pranavi, S. G. P. Dharshini, "Emotions Based Voice Supportive Model Using SVM" In *2021 International Conference on Engineering and Emerging Technologies (ICEET) IEEE*, 2021, pp. 1-4.

[6] A. Poulose, C. S. Reddy, J. H. Kim, D. S. Han, "Foreground Extraction Based Facial Emotion Recognition Using Deep Learning Xception Model". In *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), IEEE*, 2021, pp. 356-360.

[7] D. Mehta, M. F. H Siddiqui, A. Y. Javaid, "Facial emotion recognition: A survey and real-world user experiences in mixed reality", *Sensors*, 2018, vol. 18, no. 2, pp. 416.

[8] W. Mellouk, W. Handouzi, "Facial emotion recognition using deep learning: review and insights", *Procedia Computer Science*, 2020, no. 175, pp. 689-694

[9] L. Lusquino-Filho, F. M. França, P. M. Lima, "Near-optimal facial emotion classification using a WiSARD-based weightless system", In *ESANN* 2020.

[10] A. Alreshidi, M. Ullah, "Facial emotion recognition using hybrid features", *Informatics, Multidisciplinary Digital Publishing Institute*, 2020, vol. 7, no. 1, p. 6.

[11] F. Ma, B. Sun, S. Li, "Robust facial expression recognition with convolutional visual transformers", 2021 arXiv preprint arXiv:2103.16854

[12] M. Behzad, X. Li, G. Zhao, "Disentangling 3D/4D Facial Affect Recognition with Faster Multi-View Transformer" *IEEE Signal Processing Letters*, 2021, no. 28, pp. 1913-1917.

[13] H. Li, M. Sui, F. Zhao, Z. Zha, F. Wu, "MViT: Mask Vision Transformer for Facial Expression Recognition in the wild" 2021. *arXiv*, preprint arXiv:2106.04520.

[14] F. Xue, Q. Wang, G. Guo "TransFER: Learning Relation-aware Facial Expression Representations with Transformers", In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3601-3610.

[15] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, R. Seguier, R. "Learning vision transformer with squeeze and excitation for facial expression recognition", 2021, *arXiv*, preprint arXiv:2107.03107.

[16] H. Li, M. Sui, Z. Zhu, F. Zhao, "MFEViT: A Robust Lightweight Transformer-based Network for Multimodal 2D+ 3D Facial Expression Recognition", 2021, *arXiv*, preprint arXiv:2109.13086.

[17] O, Arriaga, M.Valdenegro-Toro, P. Plöger, "Real-time convolutional neural networks for emotion and gender classification", 2017, *arXiv*, preprint arXiv:1710.07557

[18] H. Yar, T. Jan, A. Hussain, S.U. Din, "Real-Time Facial Emotion Recognition and Gender Classification for Human-Robot Interaction Using CNN" The *5th International Conference on Next Generation Computing*, 2020.

[19] T. Raksarikorn, T. Kangkachit, "Facial expression classification using deep extreme inception networks" In *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, no. 06, 2018, pp. 1-5.

[20] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, S. B. Musa, S. B "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi", In *2020 Fifth International Conference on Informatics and Computing (ICIC)*, no. 11, 2020, pp. 1-9.

[21] J.H. Shah, M. Sharif, M. Yasmin, S. L. Fernandes, "Facial expressions classification and false label reduction using LDA and threefold SVM", *Pattern Recognition Letters*, 2020, no. 139, pp. 166-173.

[22] S. Lysenko, N. Seethapathi, L. Prosser, K. Kording, M.J. Johnson, "Towards Automated Emotion Classification of Atypically and Typically Developing Infants", In *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, no. 11, 2020, pp. 503-508.

[23] A. Nasuha, F. Arifin, A.S. Priambodo, N. Setiawan, N. Ahwan, "Real-Time Emotion Classification Based on Convolution Neural Network and Facial Feature", *Journal of Physics: Conference Series*, 2021, vol. 1737, no. 1,

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, T., N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale", 2021, *arXiv*, preprint arXiv:2010.11929.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, 2017, pp. 5998-6008.

Científica, vol. 26, núm. 2, pp. 01-09, julio-diciembre 2022, ISSN 2594-2921, Instituto Politécnico Nacional MÉXICO
DOI: https://doi.org/10.46842/ipn.cien.v26n2a02