EasyChair Preprint

№ 15458

Similarity-Based Positional Encoding for Enhanced Classification in Medical Images

Giorgio Leonardi, Luigi Portinale and Andrea Santomauro

November 24, 2024

Similarity-based positional encoding for enhanced classification in medical images

Giorgio Leonardi[1], Luigi Portinale[1] and Andrea Santomauro[1,*,†]

[1]Computer Science Institute, DiSIT, Università del Piemonte Orientale, V.le T. Michel 11, Alessandria, 15121, Italy

Abstract

This paper introduces a novel similarity-based positional encoding method aimed at improving the classification of medical images using Vision Transformers (ViTs). Traditional positional encoding methods focus primarily on spatial information, but they may not adequately capture the complex geometric patterns characteristic of medical images. To address this, we propose a method that utilizes convolution operations to extract geometric features, followed by a similarity matrix based on cosine similarity between image patches. This encoding is then incorporated into the ViT model, enabling it to learn more meaningful relationships beyond basic spatial positioning. The effectiveness of this method is shown through experiments on six medical imaging datasets from MedMNIST, where our approach consistently outperforms the conventional learned positional encoding. This is particularly true in datasets with prominent geometric structures like PneumoniaMNIST and BloodMNIST. The results indicate that similarity-based encoding can significantly enhance medical image classification accuracy.

Keywords

Medical Image Classification, Positional Encoding, Vision Transfomer

1. Introduction

Vision Transformers (ViTs) have revolutionized the field of computer vision, achieving state-of-the-art performance on various tasks [1]. A crucial component of these models is the positional encoding, which provides spatial information to the otherwise position-agnostic self-attention mechanism [2]. While initially designed to encode absolute or relative positions of image patches, recent studies suggest that learned positional encodings in ViTs may be capturing more than just spatial locations [3][4][5][6].

In this paper, we propose a novel perspective on learned positional encodings in Vision Transformers. We argue that these encodings are not merely learning "positions" in the traditional sense, but rather capturing high-level relationships and patterns within the visual data. This insight challenges the conventional understanding of positional encodings and opens up new avenues for improving ViT architectures. In particular, by considering complex medical images such as X-rays, histological or dermatological ones, we can notice that particular diseases or findings are usually visible in form of geometric patterns, which can be properly extracted using convolutions.

Building on this observation, we introduce a new encoding method based on similarity measures between image patches, after applying convolution operations. Our approach leverages the inherent structure and relationships within visual data, allowing the model to capture more meaningful representations that go beyond simple spatial positions. We show that this similarity-based encoding leads to improved performance and generalization across various computer vision tasks. Our contributions are threefold:

- we propose a novel similarity-based encoding method that explicitly models relationships between image patches.
- we use convolutional layers to extract geometric information from the medical images.

3rd AIxIA Workshop on Artificial Intelligence For Healthcare, 25-28 November 2024, Bolzano, Italy

*Corresponding author.

† PhD student enrolled in the National PhD in Artificial Intelligence for Health and Life Sciences, XXXVII cycle (Università Campus Bio-Medico di Roma

giorgio.leonardi@uniupo.it (G. Leonardi); luigi.portinale@uniupo.it (L. Portinale); andrea.santomauro@uniupo.it (A. Santomauro)

- we show the effectiveness of our approach through extensive experiments on medical benchmark datasets, showing improved performance compared to traditional positional encodings.

This work not only advances our understanding of Vision Transformers but also paves the way for more efficient and effective visual representation learning in deep neural networks applied to medical images interpretation.

2. Related works on positional encoding

Positional encoding is a key component for Transfomer model, introduced by [2] with the purpose of introducing spatial information to the model, which is inherently permutation-invariant. There exist four main types of different positional encoding:

- *fixed positional encoding*: it consists in fixed encodings that rely on predefined functions; for example, [2] uses sine and cosine functions for the positional encoding and connects the coding information of different frequencies to form the final positional encoding.
- *learned positional encoding*: introduced by [1], it is the most widely used and it consists in a random matrix of fixed dimensions combined with the patches, and optimized during a training phase through gradient descend.
- *learned relative positional encoding (RPE)*: introduced by [7], it focuses on the relative distances between tokens, providing more flexible and adaptive encoding. In vision transformers, RPE can help to better capture spatial correlations and context, especially for tasks like image segmentation, where the relative positions between pixels are critical.
- *task-specific positional encoding*: for specific application task (such as medical image analysis) standard encoding techniques may lack in capturing crucial contextual information. In fact, for medical image classification and segmentation some proposed methods have shown superiority over classical methods. In [8], the authors introduce a positional encoding that incorporates volumetric tokens from 3D medical images such as CT and MRI scans, focusing on enhancing segmentation through long-range contextual learning. Tang et al. [9] uses a hierarchical positional encoding adapted for volumetric medical images. The hierarchical structure improves the transformer's ability to capture both local and global features for medical image segmentation. Wang et al. [10] proposes 3D inductive positional embeddings (3D IPE) that encode both relational and absolute position information for 3D medical images. This encoding method allows the transformer to capture context effectively in 3D segmentation tasks. Yu & Triesch [11] proposes Circle Relationship Embedding (CRE) that simplifies positional encoding by utilizing the spatial arrangement of image patches in a circular manner, improving the transformer's performance on medical images.

3. Methods

The Visual Transformer (ViT) [1] is a neural network architecture designed for computer vision tasks. Figure 1 shows the ViT architecture. Unlike traditional CNNs, which process images using convolutional layers, the Visual Transformer pre-processes input images through the following steps:

- **patch extraction**: the input image is divided into a grid of non-overlapping patches. Each patch is typically a small square region of the image. For example, for an image of size 224x224 pixels, using a patch size of 16x16 generates 196 patches (14x14 grid).
- **flattening and linear projection**: each patch is then flattened into a one-dimensional vector. This means that the spatial information within each patch is encoded into a linear sequence of values.

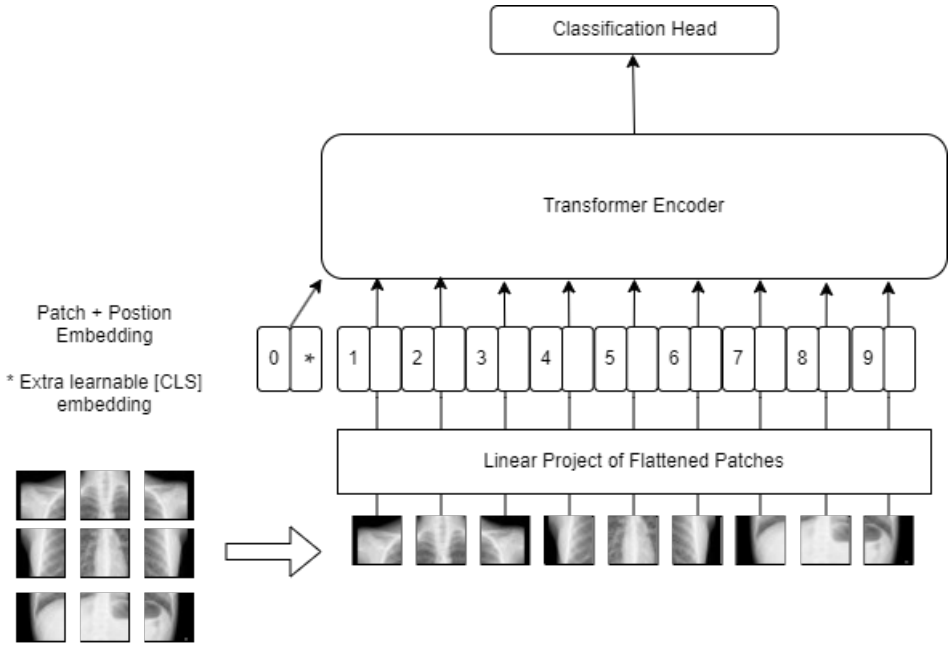

Figure 1: Vision Transformer architecture

An extra learnable parameter is preposed to the features vector, called class token; both the image patch projection and the class token have the same dimensionality.

The transformer encoder block in Vision Transformers (ViTs) is a fundamental component responsible for processing input patches and capturing global dependencies within the image. It consists of several layers, each composed of two main sub-modules: the multi-head self-attention mechanism and the position-wise feed-forward neural network. In the multi-head self-attention mechanism, the input patches are transformed into query, key, and value representations. These representations are linearly projected to multiple attention heads, which independently compute attention scores capturing the relationships between different patches. The attention scores are then weighted and combined to generate context-aware representations for each patch. Simultaneously, the position-wise feed-forward neural network applies a non-linear transformation to each patch's representation, independently. This transformation enhances the model's ability to capture complex patterns and features within the input patches.

After processing through the attention mechanism and the feed-forward network, the output representations of the patches are passed through residual connections and layer normalization, facilitating stable training and improved gradient flow. Finally, the output of the transformer encoder blocks is fed into subsequent layers or used directly for downstream tasks such as image classification. In the context of image classification, our architecture uses the class token which flows in input to a classification head, using Cross entropy loss as optimization function:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

Where N in the number of samples, C is the number of classes, y_{ij} is the true label of the i-th sample for the j-th class (either 0 or 1) and $\hat{y}_{ij}$ is the predicted probability that the i-th sample belongs to the j-th class.

3.1. Positional Encoding

One of the crucial concept in ViTs is the **positional encoding**. The positional encoding was introduced in [2] to provide spatial or sequential information to the model; in fact, self-attention mechanisms are

inherently permutation-invariant, meaning they treat input elements as a set, without any inherent order. However, both text and images have a logical ordering between words and patches, respectively. Positional encoding is used to mantain this ordering between elements: in ViT it helps the model understanding the 2D spatial relationships between image patches. This is crucial because the image is split into patches and flattened before being processed.

In section 2 we discussed the state-of-the-art for positional encoding. While initially designed to encode absolute or relative positions of image patches, recent studies suggest that learned positional encodings in ViTs may be capturing more than just spatial locations and outperform the fixed positional encoding techniques. In the next section we propose a novel similarity-based encoding method that explicitly models relationships between image patches through convolution.

4. Proposed method

Positional encodings are a class of methods which give spatial or sequential information to the self-attention mechanisms used by Transfomer models. Positional encodings are broadly studied in the literature (see Section 2) for standard Transformer model (i.e. Transformer for natural language processing). While there are few proposals for ViTs' positional encoding, the most of them adopt learned positional encoding, which seems to outperform the other techniques.

Learned positional encoding allows the model to implicitly capture the relationships between patches in a flexible way, adapting the encoding to the task at hand and discovering representations that go beyond explicit spatial coordinates, such as semantic relationships and interactions that are spatially informed. The learned positional encoding aims to capture how different parts of the image contribute to the overall context, learning patterns such as "closeness" in pixel space or object part arrangements. This can be crucial for tasks like classification and segmentation.

In medical image analysis, however, geometrical patterns (such as shapes, boundaries, and specific anatomical structures) often play a more crucial role than the general patch-to-patch correlation captured by Vision Transformers (ViT), which are unable to manage these geometrical patterns. Based on this observation, we propose a new method for positional encoding, based on convolution operations and cosine similarity between extracted features. The overall ViT architecture has not to be changed and one can exploit pre-trained models for specific task and fine-tune just the positional encoding blocks.

Let I represent the input image, divided into N patches; we embed each patch and we denote by $X \in \mathbb{R}^{N \times d}$ the matrix representing the patch embeddings extracted from the image I (N is the number of patches and d is the dimensionality of each patch embedding). Let $F_{map} \in \mathbb{R}^{N \times d}$ be the feature map computed by convolutions on the image I. The goal is to compute a similarity matrix based on the features map, and to use this similarity matrix as a positional encoding.

The cosine similarity between two vectors x_i and x_j, representing the $i-th$ and $j-th$ patch embeddings, is given by:

$$S_{ij} = \frac{x_i \cdot x_j}{||x_i|| \cdot ||x_j||} \tag{2}$$

where S_{ij} s the cosine similarity between feature i and patch j, $x_i \cdot x_j$ represents the dot product between the embeddings of patches i and j and $||x_i||$ and $||x_j||$ are the Euclidean norms (magnitudes) of the vectors x_i and x_j. The complete similarity matrix $S \in \mathbb{R}^{N \times N}$ for all features is computed as:

$$S = \frac{F_{map} \cdot F_{map}^T}{||F_{map}||_2} \tag{3}$$

where F_{map}^T is the transpose of F_{map} and $||F_{map}||_2$ normalizes each row of the features matrix to compute the cosine similarities row-wise. Next, we apply a linear transformation to the similarity matrix to map it to the same dimensionality as the original patch embeddings. Let's define this transformation using a learnable weight matrix $W \in \mathbb{R}^{N \times d}$:

$$PE = SW \tag{4}$$

where $PE \in \mathbb{R}^{N \times d}$ is the positional encoding matrix based on similarity between patches, $S \in \mathbb{R}^{N \times N}$ is the cosine similarity matrix and $W \in \mathbb{R}^{N \times d}$ is a learnable projection matrix that transforms the similarity information to match the dimensionality d of the patch embeddings.

Since $X \in \mathbb{R}^{N \times d}$ represents the patch embeddings, the positional ecoding PE can be added to X to produce the final input to the transformer:

$$X' = X + PE \tag{5}$$

where $X' \in \mathbb{R}^{N \cdot d}$ is the modified patch embedding, which now includes the similarity-based positional encoding.

The learned convolutional filters extract geometrical information from the input images, which are essentials in medical image analysis. Through the usage of these geometrical features as encoding, we help the Transfomer architecture to learn patterns which can be useful for the specific task.

5. Experiments and Results

In order to compare learned positional encoding (which is the state-of-the-art encoding for ViTs) with our proposed encoding based on similarity, we have defined an architecture with fixed hyperparameters (e.g number of transfomer layer, number of heads, embedding dimension, etc.) and we have trained the network on specific medical datasets with both encoding techniques, by comparing the results.

5.1. Datasets

We resort to 6 different datasets which are part of MedMNIST[12][13], a collection of standardized, lightweight, and preprocessed biomedical image datasets designed for evaluating machine learning algorithms, especially in the medical field. MedMNIST includes several image classification tasks, primarily focusing on diverse types of medical data such as internal pathology, dermatology, ophthalmology, and more. It is inspired by the well-known MNIST dataset (a handwritten digit classification benchmark), but tailored for biomedical images. The images in MedMNIST can be downloaded in different size,and we choose a $224x224$ format to remain consistent with the standard input size of the ViT introduced in [2]. Specifically we used:

- **PneumoniaMNIST**, a dataset of chest X-ray images, specifically designed to detect and classify pneumonia. The dataset was created by extracting data from the NIH's ChestX-ray14 database, which contains X-rays labeled with various pulmonary conditions. Task: binary classification.
- **PathMNIST**, a dataset based on pathology images of breast cancer tissue slides, collected from the CAMELYON16 challenge dataset. The objective is to classify different subtypes of breast cancer from microscopy images. Task: multi-class classification.
- **DermaMNIST**, consisting of dermatology images aimed at classifying different types of skin lesions. This dataset is derived from the HAM10000 dataset, which is commonly used for the classification of skin diseases, such as melanomas and benign lesions. Task: multi-class classification.
- **BreastMNIST**, a dataset derived from ultrasound images for the classification of breast cancer. The images represent different conditions, including benign, malignant, and normal tissue. Task: binary or multi-class classification.
- **TissueMNIST**, a dataset that focuses on histological images from human tissue samples. It contains tissue images from various organs, extracted from the HuBMAP dataset. Each image represents a section of tissue that can be classified into different cell types. Task: multi-class classification.
- **BloodMNIST**, a dataset of blood smear images for the classification of blood cells. It is derived from the Atlas of Blood dataset and contains images of different blood cell types, which is useful for tasks related to hematology. Task: multi-class classification.

(a) BloodMNIST (b) PneumoniaMNIST (c) PathMNIST

Figure 2: Sample images from BloodMNIST, PneumoniaMNIST, and PathMNIST datasets

Encoding type	Dataset	Accuracy
Learned PE	PneumoniaMNIST	85.89%
Similarity PE	PneumoniaMNIST	**91.18**%
Learned PE	PathMNIST	83.14%
Similarity PE	PathMNIST	**88.53**%
Learned PE	DermaMNIST	70.97%
Similarity PE	DermaMNIST	**73.61**%
Learned PE	BreastMNIST	73.01%
Similarity PE	BreastMNIST	**76.97**%
Learned PE	TissueMNIST	50.87%
Similarity PE	TissueMNIST	**54.81**%
Learned PE	BloodMNIST	83.95%
Similarity PE	BloodMNIST	**94.18**%

Table 1
Results on 6 benchmark datasets and comparison between Learned Positional Encoding and Similarity Positional Encoding

5.2. Results

In this section we compare the results we have obtained by considering both standard learned positional encoding and similarity positional encoding. Table 1 shows such results in term of accuracy. It is possible to see that the proposed similarity positional encoding brings to better accuracy for every benchmark dataset we have used.

For BloodMNIST, PneumoniaMNIST and PathMNIST the accuracy gap between our positional encoding and learned positional encoding is bigger than for the other datasets. This may be explained by the fact that these datasets contain more geometrical structure than the other, and since our positional encoding is designed to add to the model the ability to recognize geometric shapes, the behaviour is consistent with the hypotheses. Figure 2 shows an example of BloodMNIST's image (2a), PneumoniaMNIST's image (2b) and PathMNIST's image (2c). In contrast, datasets as DermaMNIST, in which there are no such geometrical features, the resulting performance are slightly increased but overall comparable.

6. Conclusion and future works

In this work we have proposed a new technique for positional encoding in medical image analysis. Starting from the rationale of having geometrical structures in medical images, we have developed a new encoding technique that employs convolutions to extract geometrical information from the

input images and compute cosine similarity between extracted features, in such a way to use this information as positional encoding. We have shown that, in terms of accuracy, this new positional encoding outperforms the standard learned positional encoding, especially in images featuring geometric structures.

As future works we want to focus on two different tasks:

- compare the attention masks generated by the two different encodings, in order to evaluate if the explainability of the model increases introducing geometrical features;
- compare our positional encoding with other proposed method, listed in section 2

Acknowledgments

Andrea Santomauro's PhD research is co-financed by ARLANIS REPLY. We also acknowledge the use of the Chameleon Cloud testbed (`https://chameleoncloud.org/`) that has allowed us to perform all the experiments described in the present work.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[3] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, 2023. URL: https://arxiv.org/abs/2104.09864. arXiv:2104.09864.

[4] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[5] O. Press, N. A. Smith, M. Lewis, Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL: https://arxiv.org/abs/2108.12409. arXiv:2108.12409.

[6] P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, 2018. URL: https://arxiv.org/abs/1803.02155. arXiv:1803.02155.

[7] K. Wu, H. Peng, M. Chen, J. Fu, H. Chao, Rethinking and improving relative position encoding for vision transformer, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10013–10021. doi:10.1109/ICCV48922.2021.00988.

[8] A. Hatamizadeh, Z. Xu, D. Yang, W. Li, H. Roth, D. Xu, Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation, 2022. URL: https://arxiv.org/abs/2204.00631. arXiv:2204.00631.

[9] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, A. Hatamizadeh, Self-supervised pre-training of swin transformers for 3d medical image analysis, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20698–20708. doi:10.1109/CVPR52688.2022.02007.

[10] L. Wang, X. Wang, B. Zhang, X. Huang, C. Bai, M. Xia, P. Sun, Multi-scale hierarchical transformer structure for 3d medical image segmentation, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 1542–1545. doi:10.1109/BIBM52615.2021.9669799.

[11] Z. Yu, J. Triesch, Cre: Circle relationship embedding of patches in vision transformer, ESANN 2023 proceesdings (2023). doi:10.14428/esann/2023.es2023-75.

[12] J. Yang, R. Shi, B. Ni, Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis, in: IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 191–195.

[13] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification, Scientific Data 10 (2023) 41.