ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200447

TransSketchNet: Attention-Based Sketch Recognition Using Transformers

Gaurav Jain[1,2] and **Shivang Chopra**[1,2] and **Suransh Chopra**[1,2] and **Anil Singh Parihar**[2]

Abstract. Sketches have been employed since the ancient era of cave paintings for simple illustrations to represent real-world entities. The abstract nature and varied artistic styling makes automatic recognition of drawings more challenging than other areas of image classification. Moreover, the representation of sketches as a sequence of strokes instead of raster images introduces them at the correct abstract level. However, dealing with images as a sequence of small information makes it challenging. In this paper, we propose a Transformer-based network, dubbed as TransSketchNet, for sketch recognition. This architecture incorporates ordinal information to perform the classification task in real-time through vector images.

1 Introduction

The ability to draw and comprehend eclectic notions through sketches reflects the intellectual capabilities of human beings. Drawing has long been a part of human behavior, mutating its utility from rock carvings in the ancient era, to blueprints of drafts in the modern age. In this work, we focus on *sketch recognition*, which aims to identify human drawings and classify them into their respective categories.

Perceiving sketch is a challenging task due to two main reasons, (1) heterogeneous representations and (2) level of abstraction. Representation of the same object is dependent upon the interpretation of that object, which consequently leads to high intra-class variation between samples of the same class. Fig. 1 illustrates this diversity among different samples of the same class. For example, distinct views of a bat in Fig. 1 (a) include different representations which render the task of finding patterns relatively difficult for the models. Alternatively, sketches of one class could be interpreted as belonging to another class. For instance, the second sketch of the sea turtle class in Fig. 1 (b) resembles an ant shown in Fig. 1 (c). Hence, learning such overlapping representations which may exhibit a significantly anomalous, yet acceptable, view proves to be an arduous task.

In order to address the aforementioned issues, a transformer-based approach is proposed for attentive sketch recognition. This is the first approach to the best of our knowledge that employs transformers for sketch recognition. We leverage the attention mechanism of Transformers to identify objects using the vector image format. For this, sketches are interpreted as a sequence of strokes, like humans actually comprehend drawings in real-time.

Figure 1. Visualization of few classes in the QuickDraw [2] dataset, (a) Bat, (b) Sea Turtle, (c) Ant, (d) Star. Attention heatmaps depicting the relative importance of strokes while inference.

2 Proposed Methodology

Input Pre-processing: The input data is pre-processed to transform the sketches into the vector-image format. In this, $\mathcal{S}_v$ is a sketch with sequence of strokes s_i, i.e. $\mathcal{S}_v = \{s_1, s_2, ..., s_n\}$, where n is the sequence length. Each stroke, s_i, is defined using a 3-point format:

$$s_i = \{\Delta x_i, \Delta y_i, p_i\}, \forall i \in \{1, 2, ..., n\} \quad (1)$$

where $(\Delta x_i, \Delta y_i)$ is the offset distance in the x and y direction. For each sketch, $(\Delta x_1, \Delta y_1)$ begin with origin as the initial starting point. Pen-state, p_i is a binary variable, with $p_i = 1$ indicating that the pen is in contact with surface, while $p_i = 0$ represents that the pen is lifted off the surface and moved from the previous point in the direction of offsets.

Architecture: Fig. 2 illustrates the proposed architecture, which consists of two modules, (1) *auto-encoder*, and (2) *transformer-encoder*. The *auto-encoder* performs two major functions, (1) extract latent vector, $\mathcal{Z}_v$, and (2) reshape the input dimension to feed a larger context vector to the transformer module effectively. The input $\mathcal{S}_v$ is not directly fed into the auto-encoder. Instead, it is decomposed as $\mathcal{S}_v = \{\mathcal{S}_v^{\Delta}, \mathcal{S}_v^{p}\}$, where $\mathcal{S}_v^{\Delta} = (\Delta x_i, \Delta y_i)$ and $\mathcal{S}_v^{p} = p_i$. We only feed the offsets, $\mathcal{S}_v^{\Delta} \in \mathbb{R}^{2\times n}$, to obtain latent vector, $\mathcal{Z}_v \in \mathbb{R}^{127\times n}$, with a sequence length of n. To preserve the temporal nature of our data, we use *coordinate embedding*, which computes a tuple with (position, time). In the *transformer module*, we use 10 identical transformer blocks connected via skip-connections. Each block consists of multi-head self-attention with $m = 8$ heads. The custom *Extract layer* is used to pick the features (ext_{out}) corresponding

[1] Equal Contribution.
[2] Machine Learning Research Laboratory, Department of Computer Science & Engineering, Delhi Technological University, New Delhi 110042, India. email: {gauravjain13298, shivangchopra11, suransh2008, parihar.anil}@gmail.com

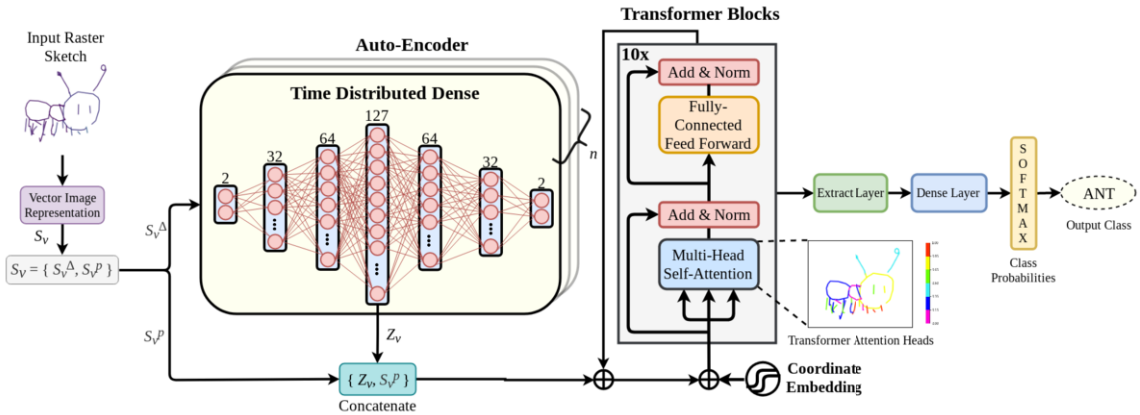

Figure 2. Overview of the proposed Network depicting the pre-processing of raster sketch to form the vector image representation. Projecting the input to a higher latent dimension using the Auto-Encoder and the final classification using the attention maps from the Transformer blocks.

to the fist timestamp of the transformer output. Further, ext_{out} is passed through a dense layer with a softmax activation function to obtain class probabilities $\Psi = \{\psi_1, \psi_2, ..., \psi_c\}$, where c represents the number of classes. Finally, dense layers with softmax activation facilitate the model to return class probabilities.

Method	5	20	50
HOG-SVM [1]	75.21%	66.79%	63.22%
Fisher-Vectors [5]	79.53%	75.80%	72.90%
AlexNet [4]	77.18%	75.22%	73.06%
Sketch-a-Net v2 [6]	94.78%	88.64%	85.19%
Resnet50-CNN [3]	**96.47%**	90.06%	86.20%
TransSketchNet (Ours)	96.21%	**90.31%**	**88.72%**

Table 1. Comparative evaluation of recognition accuracy on the Quick Draw dataset, with (a) 5 classes, (b) 20 classes, (c) 50 classes. Values in **bold** depict the best accuracy for each subset.

3 Results and Evaluation

The performance of the proposed Transformer-based network is evaluated using the benchmark Quick Draw dataset [2]. Further, attention heatmaps have been plotted to support the use of transformers for sketch recognition. A train-test split of 80-20% is performed with $50,000$ samples per class for training, and $10,000$ samples per class during testing.

Comparative Analysis. We compared the proposed TransSketchNet with three types of approaches, (1) traditional classifiers, (2) CNN-based approaches, and (3) RNN-based approaches. Table 1 reports the recognition accuracy. TransSketchNet outperforms the state-of-the-art approaches for 20 and 50 classes. However, Resnet50-CNN [6] performs marginally better than TransSketchNet with a gain of 0.26% for dataset with 5 classes. For the proposed approach, increasing the number of classes observes a lower drop in accuracy, compared to other approaches. Due to limited resource availability, analysis over the complete dataset could not be performed. To account for this, we randomly selected c classes, where $c \in \{5, 20, 50\}$, and report the averages over 5-fold cross-validation. Quantitative evaluation confirms the effectiveness of representing sketches as vector images. Moreover, this representation facilitates real-time recognition of sketches.

Attention Analysis. The proposed method introduces transformers to exploit the sequence of strokes to draw attention to parts of sketches. Fig. 1 shows few attention heatmaps. For the *bat* class, attention on both the wings are equally focused. This is analogous to how humans recognize bats through their peculiar wing shape. It is interesting to note that for the *star* class, equally high attention is given to the complete structure. Perhaps, structures like stars that are symmetric in nature and constitute a distinct shape are identified based on the overall view of the sketches. Similarly, certain parts of the sketch are more important in recognizing an object; which is supported by the attention heatmaps presented.

4 Conclusion and Future Work

In this work, we proposed TransSketchNet for recognizing sketches using vector images. The autoencoder extracts information from the input data in the form of a latent vector. This renders the input dimension compatibility between the input data and transformer-blocks. Further, transformer-blocks enable the model to focus on characteristic information from each sketch. The proposed approach achieves favorable recognition accuracy when compared with state-of-the-art approaches. In the future, Transformers can be adapted to solve challenging problems in the domain of computer vision.

REFERENCES

[1] Mathias Eitz, James Hays, and Marc Alexa, 'How do humans sketch objects?', *ACM Trans. Graph.*, **31**(4), 44–1, (2012).
[2] David Ha and Douglas Eck, 'A neural representation of sketch drawings', *CoRR*, **abs/1704.03477**, (2017).
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).
[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks', in *Advances in neural information processing systems*, pp. 1097–1105, (2012).
[5] Rosália G Schneider and Tinne Tuytelaars, 'Sketch classification and classification-driven analysis using fisher vectors', *ACM Transactions on Graphics (TOG)*, **33**(6), 174, (2014).
[6] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales, 'Sketch-a-net: A deep neural network that beats humans', *International journal of computer vision*, **122**(3), 411–425, (2017).