

# Multi-Graph Transformer for Free-Hand Sketch Recognition

Peng Xu<sup>1</sup> Chaitanya K. Joshi<sup>1</sup> Xavier Bresson<sup>1</sup>

## Abstract

Learning meaningful representations of free-hand sketches remains a challenging task given the signal sparsity and the high-level abstraction of sketches. Existing techniques have focused on exploiting either the static nature of sketches with Convolutional Neural Networks (CNNs) or the temporal sequential property with Recurrent Neural Networks (RNNs). In this work, we propose a new representation of sketches as multiple sparsely connected graphs. We design a novel Graph Neural Network (GNN), the Multi-Graph Transformer (MGT), for learning representations of sketches from multiple graphs which simultaneously capture global and local geometric stroke structures, as well as temporal information. We report extensive numerical experiments on a sketch recognition task to demonstrate the performance of the proposed approach. Particularly, MGT applied on 414k sketches from Google QuickDraw: (i) achieves small recognition gap to the CNN-based performance upper bound (72.80% vs. 74.22%), and (ii) outperforms all RNN-based models by a significant margin. To the best of our knowledge, this is the first work proposing to represent sketches as graphs and apply GNNs for sketch recognition. Code and trained models are available at [https://github.com/PengBoXiangShang/multigraph\\_transformer](https://github.com/PengBoXiangShang/multigraph_transformer).

## 1. Introduction

Free-hand sketches are drawings made without the use of any instruments. Sketches are different from traditional images: they are formed of temporal sequences of strokes (Ha & Eck, 2018; Xu et al., 2018), while images are static collections of pixels with dense color and texture patterns. Sketches capture high-level abstraction of visual objects

<sup>1</sup>School of Computer Science and Engineering, Nanyang Technological University, Singapore. Correspondence to: Peng Xu <peng.xu@ntu.edu.sg>.

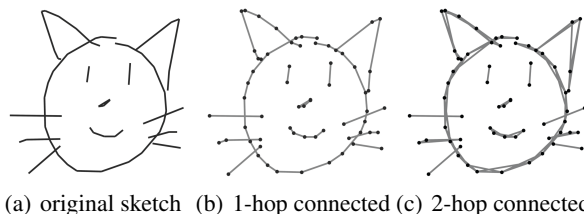


Figure 1. Sketches can be seen as sets of curves and strokes, which are discretized by graphs.

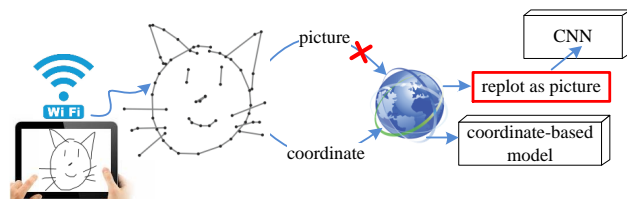


Figure 2. In sketch-based human-computer interaction scenarios, it is time-consuming to render and transfer pictures of sketches. Solely transferring stroke coordinates leads to real-time applications.

with very sparse information compared to regular images, which makes the modelling of sketches unique and challenging.

The modern prevalence of touchscreen devices has led to a flourishing of sketch-related applications in recent years, including sketch recognition (Liu et al., 2019), sketch scene understanding (Ye et al., 2016), sketch hashing (Xu et al., 2018), sketch-based image retrieval (Sangkloy et al., 2016; Liu et al., 2017; Shen et al., 2018; Collomosse et al., 2019; Dutta & Akata, 2019; Dey et al., 2019), and sketch-related generative models (Ha & Eck, 2018; Chen & Hays, 2018; Lu et al., 2018; Liu et al., 2019).

If we assume sketches to be 2D static images, CNNs can be directly applied to sketches, such as “Sketch-a-Net” (Yu et al., 2015). If we now suppose that sketches are ordered sequences of point coordinates, then RNNs can be used to recursively capture the temporal information, *e.g.*, “SketchRNN” (Ha & Eck, 2018).

In this work, we introduce a new representation of sketches with *graphs*. We assume that sketches are sets of curves

and strokes, which are discretized by a set of points representing the graph nodes. This view offers high flexibility to encode different sketch geometric properties as we can decide different connectivity structures between the node points. We use two types of graphs to represent sketches: intra-stroke graphs and extra-stroke graphs. The first graphs capture the local geometry of strokes, independently to each other, with for example 1-hop or 2-hop connected graphs, see Figure 1. The second graphs encode the global geometry and temporal information of strokes. Another advantage of using graphs is the freedom to choose the node features. For sketches, spatial, temporal and semantic information is available with the stroke point coordinates, the ordering of points, and the pen state information, respectively. In summary, representing sketches with graphs offers a universal representation that can make use of global and local spatial sketch structures, as well as temporal and semantic information.

To exploit these graph structures, we propose a new Transformer (Vaswani et al., 2017) architecture that can use multiple sparsely connected graphs. It is worth reporting that a direct application of the original Transformer model on the input spatio-temporal features provides poor results. We argue that the issue comes from the graph structure in the original Transformer which is a fully connected graph. Although fully-connected word graphs work impressively for Natural Language Processing, where the underlying word representations themselves contain rich information, such dense graph structures provide poor inductive bias (Battaglia et al., 2018) for 2D sketch tasks. Transformers require sketch-specific design coming from geometric structures. This led us to naturally extend Transformers to multiple arbitrary graph structures. Moreover, graphs provide more robustness to handle noisy and style-changing sketches as they focus on the geometry of strokes and not on the specific distribution of points.

Another advantage of using domain-specific graphs is to leverage the sparsity property of discretized sketches. Observe that intra-stroke and extra-stroke graphs are *highly sparse* adjacency matrices. In practical sketch-based human-computer interaction scenarios, it is time-consuming to directly transfer the original sketch picture from user touch-screen devices to the back-end servers. To ensure real-time applications, transferring the stroke coordinates as a character string would be more beneficial, see Figure 2.

Our main contributions can be summarised as follows:

- (i) We propose to model sketches as sparsely connected graphs, which are flexible to encode local and global geometric sketch structures. To the best of our knowledge, it is the first time that graphs are proposed for representing sketches.
- (ii) We introduce a novel Transformer architecture that can

handle multiple arbitrary graphs. Using intra-stroke and extra-stroke graphs, the proposed *Multi-Graph Transformer* (MGT) learns both local and global patterns along sub-components of sketches.

(iii) Numerical experiments demonstrate the performances of our model. MGT significantly outperforms RNN-based models, and achieves small recognition gap to CNN-based architectures. This is promising for real-time sketch-based human-computer interaction systems. Note that CNNs are our performance upper bound for coordinate-based models that involve truncating coordinate sequences, *e.g.*, RNN or Transformer based architectures.

## 2. Related Work

**Neural Network Architectures for Sketches** CNNs are a common choice for feature extraction from sketches. “Sketch-a-Net” (Yu et al., 2015) was the first CNN-based model having a sketch-specific architecture. It was directly inspired from AlexNet (Krizhevsky et al., 2012) with larger first layer filters, no layer normalization, larger pooling sizes, and high dropout. Song et al. (2017) further improved Sketch-a-Net by adding spatial-semantic attention layers. “SketchRNN” (Ha & Eck, 2018) was a seminal work to model temporal stroke sequences with RNNs. A CNN-RNN hybrid architecture for sketches was proposed in (Sarvadevabhatla et al., 2016).

In this work, we propose a novel Graph Neural Network architecture for learning sketch representations from multiple sparse graphs, combining both stroke geometry and temporal order.

**Graph Neural Networks** Graph Neural Networks (GNNs) (Bruna et al., 2014; Defferrard et al., 2016; Sukhbaatar et al., 2016; Kipf & Welling, 2017; Hamilton et al., 2017; Monti et al., 2017) aim to generalize neural networks to non-Euclidean domains such as graphs and manifolds. GNNs iteratively build representations of graphs through recursive neighborhood aggregation (or message passing), where each graph node gathers features from its neighbors to represent local graph structure.

**Transformers** The Transformer architecture (Vaswani et al., 2017), originally proposed as a powerful and scalable alternative to RNNs, has been widely adopted in the Natural Language Processing community for tasks such as machine translation (Edunov et al., 2018; Wang et al., 2019), language modelling (Radford et al., 2018; Dai et al., 2019), and question-answering (Devlin et al., 2019; Yang et al., 2019).

Transformers for NLP can be regarded as GNNs which use self-attention (Veličković et al., 2018) for neighborhood aggregation on fully-connected word graphs (Ye et al., 2019).

However, GNNs and Transformers perform poorly when sketches are modelled as fully-connected graphs. This work advocates for the injection of inductive bias into Transformers through domain-specific graph structures.

### 3. Method

#### 3.1. Notation

We assume that the training dataset  $D$  consists of  $N$  labeled sketches:  $D = \{(\mathbf{X}_n, z_n)\}_{n=1}^N$ . Each sketch  $\mathbf{X}_n$  has a class label  $z_n$ , and can be formulated as a  $S$ -step sequence  $[\mathbf{C}_n, \mathbf{f}_n, \mathbf{p}] \in \mathbb{R}^{S \times 4}$ .  $\mathbf{C}_n = \{(x_n^s, y_n^s)\}_{s=1}^S \in \mathbb{R}^{S \times 2}$  is the coordinate sequence of the sketch points  $\mathbf{X}_n$ . All sketch point coordinates have been uniformly scaled to  $x_n^s, y_n^s \in [0, 256]^2$ . If the true length of  $\mathbf{C}_n$  is shorter than  $S$  then the vector  $[-1, -1]$  is used for padding. Flag bit vector  $\mathbf{f}_n \in \{f_1, f_2, f_3\}^{S \times 1}$  is a ternary integer vector that denotes the pen state sequence corresponding to each point of  $\mathbf{X}_n$ . It is defined as follows:  $f_1$  if the point  $(x_n^s, y_n^s)$  is a starting or ongoing point of a stroke,  $f_2$  if the point is the ending point of a stroke, and  $f_3$  for a padding point. Vector  $\mathbf{p} = [0, 1, 2, \dots, S-1]^T$  is a positional encoding vector that represents the temporal position of the points in each sketch  $\mathbf{X}_n$ .

Given  $D$ , we aim to model  $\mathbf{X}_n$  as multiple sparsely connected graphs and learn a deep embedding space, where the high-level semantic tasks can be conducted, *e.g.*, sketch recognition.

#### 3.2. Multi-Modal Input Layer

Given a sketch  $\mathbf{X}_n$ , we model its  $S$  stroke points as  $S$  nodes of a graph. Each node has three features: (i)  $\mathbf{C}_n^s$  is the spatial positional information of the current stroke point  $s$ , (ii)  $\mathbf{f}_n^s$  is the pen state of the current stroke point. This information helps to identify the stroke points belonging to the same stroke, and (iii)  $\mathbf{p}^s$  is the temporal information of the current stroke point. As sketching is a dynamic process, it is important to use the temporal information.

The complete model architecture for our Multi-Graph Transformer is presented in Figure 3. Let us start by describing the input layer. The final vector at node  $s$  of the multi-modal input layer is defined as

$$(\mathbf{h}_n^s)^{(l=0)} = \mathcal{C}(\mathcal{E}_1(\mathbf{C}_n^s), \mathcal{E}_2(\mathbf{f}_n^s), \mathcal{E}_2(\mathbf{p}^s)), \quad (1)$$

where  $\mathcal{E}_1(\mathbf{C}_n^s)$  is the embedding of  $\mathbf{C}_n^s$  with a linear layer of size  $2 \times \hat{d}$ ,  $\mathcal{E}_2(\mathbf{f}_n^s)$  and  $\mathcal{E}_2(\mathbf{p}^s)$  are the embeddings of the flag bit  $\mathbf{f}_n^s$  (3 discrete values) and the position encoding  $\mathbf{p}^s$  ( $S$  discrete values) from an embedding dictionary of size  $(S+3) \times \hat{d}$ , and  $\mathcal{C}(\cdot, \cdot)$  is the concatenation operator. The node vector  $(\mathbf{h}_n^s)^{(l=0)}$  has dimension  $d = 3\hat{d}$ . The design of the input layer was selected after extensive ablation studies, which are described in subsequent sections.

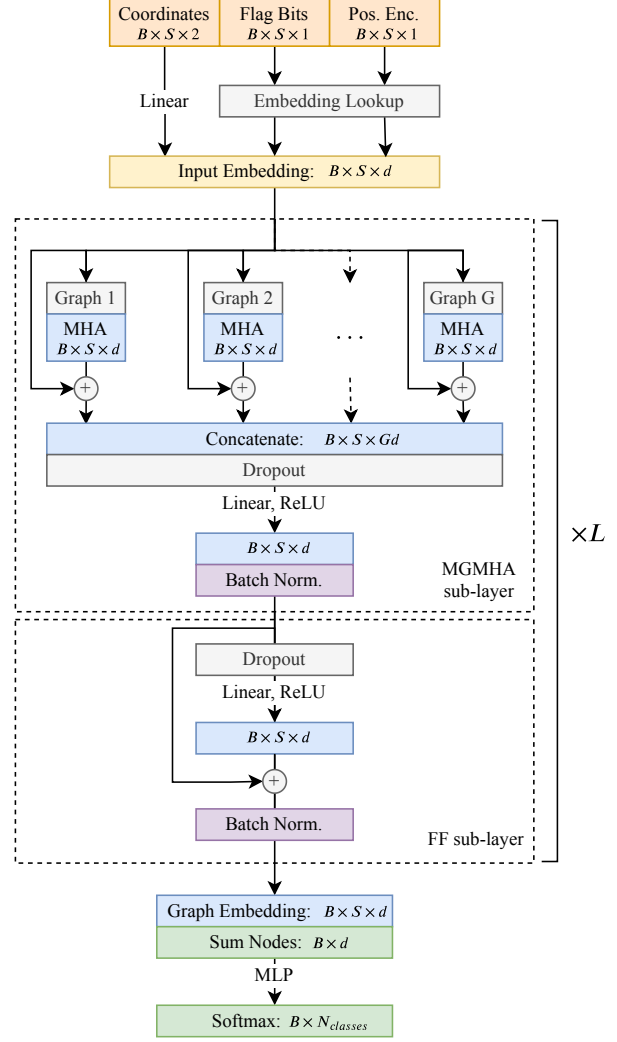


Figure 3. Multi-Graph Transformer architecture. Each MGT layer is composed of (i) a Multi-Graph Multi-Head Attention (MGMA) sub-layer and (ii) a position-wise fully connected Feed-Forward (FF) sub-layer. See details in text. “B” denotes batch size.

#### 3.3. Multi-Graph Transformer

The initial node embedding  $(\mathbf{h}_n^s)^{(l=0)}$  is updated by stacking  $L$  Multi-Graph Transformer (MGT) layers (7). Let us describe all layers.

**Graph Attention Layer** Let  $\mathbf{A}$  be a graph adjacency matrix of size  $S \times S$  and  $\mathbf{Q} \in \mathbb{R}^{S \times d_q}$ ,  $\mathbf{K} \in \mathbb{R}^{S \times d_k}$ ,  $\mathbf{V} \in \mathbb{R}^{S \times d_v}$  be the query, key, and value matrices. We define a graph attention layer as

$$\text{GraphAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{A}) = \mathbf{A} \odot \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad (2)$$

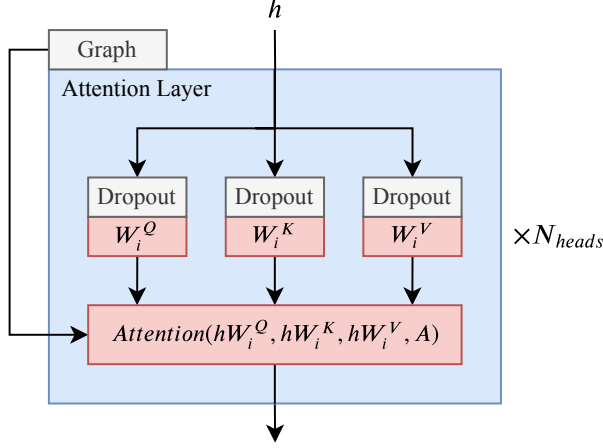


Figure 4. Multi-Head Attention Layer, consisting of several Graph Attention Layers in parallel.

where  $\odot$  is the Hadamard product. We simply weight the “Scaled Dot-Product Attention” (Vaswani et al., 2017) with the graph edge weights. We set  $d_q = d_k = d_v = \frac{d}{I}$ , where  $I$  is the number of attention heads.

**Multi-Head Attention Layer** We aggregate the graph attentions with multiple heads:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{A}) = \mathcal{C}(\text{head}_1, \dots, \text{head}_I) \mathbf{W}^O, \quad (3)$$

where  $\mathbf{W}^O \in \mathbb{R}^{I d_v \times d}$  and each attention head is computed with the graph attention layer (2):

$$\text{head}_i = \text{GraphAttention}(\mathbf{Q} \mathbf{W}_i^Q, \mathbf{K} \mathbf{W}_i^K, \mathbf{V} \mathbf{W}_i^V, \mathbf{A}), \quad (4)$$

where  $\mathbf{W}_i^Q \in \mathbb{R}^{d \times d_q}$ ,  $\mathbf{W}_i^K \in \mathbb{R}^{d \times d_k}$ , and  $\mathbf{W}_i^V \in \mathbb{R}^{d \times d_v}$ . We add dropout (Srivastava et al., 2014) before the linear projections of  $\mathbf{Q}$ ,  $\mathbf{K}$  and  $\mathbf{V}$ . An illustration of the Multi-Head Attention Layer is presented in Figure 4.

**Multi-Graph Multi-Head Attention Layer** Given a set of adjacency graph matrices  $\{\mathbf{A}_g\}_{g=1}^G$ , we can concatenate Multi-Head Attention Layers:

$$\text{MultiGraphMultiHeadAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \{\mathbf{A}_g\}_{g=1}^G) = \text{ReLU}(\mathcal{C}(\text{ghead}_1, \dots, \text{ghead}_G) \mathbf{W}^{\tilde{O}}), \quad (5)$$

where  $\mathbf{W}^{\tilde{O}} \in \mathbb{R}^{G d \times d}$  and each Multi-Head Attention Layer is computed with (3):

$$\text{ghead}_g = \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{A}_g). \quad (6)$$

**Multi-Graph Transformer Layer** The Multi-Graph

Transformer (MGT) at layer  $l$  for node  $s$  is defined as

$$\begin{aligned} (\mathbf{h}_n^s)^{(l)} &= \text{MGT}((\mathbf{h}_n^s)^{(l-1)}) \\ &= \hat{\mathbf{h}}_n^s + \text{FF}^{(l)}(\hat{\mathbf{h}}_n^s), \end{aligned} \quad (7)$$

where the intermediate feature representation  $\hat{\mathbf{h}}_n^s$  is defined as:

$$\hat{\mathbf{h}}_n^s = (\text{MGMHA}_n^s)^{(l)}((\mathbf{h}_n^1)^{(l-1)}, \dots, (\mathbf{h}_n^S)^{(l-1)}). \quad (8)$$

The MGT layer is thus composed of (i) a Multi-Graph Multi-Head Attention (MGMHA) sub-layer (5) and (ii) a position-wise fully connected Feed-Forward (FF) sub-layer. Each MHA sub-layer (6) and FF (7) has residual connections (He et al., 2016) and batch normalization (Ioffe & Szegedy, 2015). See Figure 3 for an illustration.

### 3.4. Sketch Embedding and Classification Layer

Given a sketch  $\mathbf{X}_n$  with  $t_n$  key points, its continuous representation  $\mathbf{h}_n$  is simply given by the sum over all its node features from the last MGT layer:

$$\mathbf{h}_n = \sum_{s=1}^{t_n} (\mathbf{h}_n^s)^{(L)}. \quad (9)$$

Finally, we use a Multi-Layer Perceptron (MLP) to classify the sketch representation  $\mathbf{h}_n$ , see Figure 3.

### 3.5. Sketch-Specific Graphs

In this section, we discuss the graph structures we used in our Graph Transformer layers. We considered two types of graphs, which capture local and global geometric sketch structures.

The first class of graphs focus on representing the local geometry of individual strokes. We choose  $K$ -hop graphs to describe the local geometry of strokes. The intra-stroke adjacency matrix is defined as follows:

$$\mathbf{A}_{n,ij}^{K\text{-hop}} = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{K\text{-hop}} \text{ and } j \in \text{global}(i), \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where  $\mathcal{N}_i^{K\text{-hop}}$  is the  $K$ -hop neighborhood of node  $i$  and  $\text{global}(i)$  is the stroke of node  $i$ .

The second class of graphs capture the global and temporal relationships between the strokes composing the whole sketch. We define the extra-stroke adjacency matrix as follows:

$$\mathbf{A}_{n,ij}^{\text{global}} = \begin{cases} 1 & \text{if } |i - j| = 1 \text{ and } \text{global}(i) \neq \text{global}(j), \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

This graph will force the network to pay attention between two points belonging to two distinct strokes but consecutive in time, thus allowing the model to understand the relative arrangement of strokes.



Table 1. Summary statistics for our subset of QuickDraw.

Set	# Samples	# Truncated (ratio)	# Key Points			
			max	min	mean	std
Training	345,000	11788 (3.42%)	100	2	43.26	21.85
Validation	34,500	1218 (3.53%)	100	2	43.24	21.89
Test	34,500	1235 (3.58%)	100	2	43.20	21.93

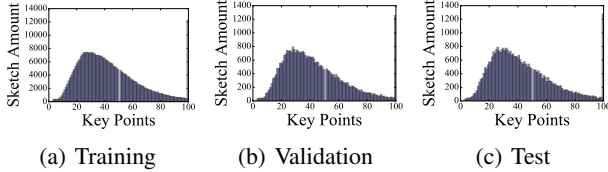


Figure 5. Histograms of key points per sketch for our subset of QuickDraw. The sharp spike at 100 key points is due to truncation.

## 4. Experiments

### 4.1. Experimental Setting

**Dataset and Pre-Processing** Google QuickDraw (Ha & Eck, 2018)<sup>1</sup> is the largest available sketch dataset containing 50 Million sketches as simplified stroke key points in temporal order, sampled using the RamerDouglasPeucker algorithm after uniformly scaling image coordinates within 0 to 256. Unlike smaller crowd-sourced sketch datasets, *e.g.*, TU-Berlin (Eitz et al., 2012), QuickDraw samples were collected via an international online game where users have only 20 seconds to sketch objects from 345 classes, such as cats, dogs, clocks, *etc.* Thus, sketch classification on QuickDraw not only involves a diversity of drawing styles, but can also be highly abstract and noisy, making it a challenging and practical test-bed for comparing the effectiveness of various neural network architectures. Following recent practices (Dey et al., 2019; Xu et al., 2018), we create random training, validation and test sets from the full dataset by sampling 1000, 100 and 100 sketches respectively from each of the 345 categories in QuickDraw. Following (Xu et al., 2018), we truncate or pad all samples to a uniform length of 100 key points/steps to facilitate efficient training of RNN and GNN-based models. We provide summary statistics for our training, validation and test sets in Table 1, and histograms visualizing the key points per sketch are shown in Figure 5.

**Evaluation Metrics** Our evaluation metric for sketch recognition is “top K accuracy”, the proportion of samples whose true class is in the top K model predictions, for values  $k = 1, 5, 10$ . (Note that  $\text{acc.}@k = 1.0$  means 100%)

**Implementation Details** For fair comparison under similar hardware conditions, all experiments were implemented in PyTorch (Paszke et al., 2019) and run on one Nvidia

1080Ti GPU. For Transformer models, we use the following hyperparameter values:  $S = 100$ ,  $L = 4$ ,  $\hat{d} = 128$ ,  $G = 3$  ( $\mathbf{A}^{1\text{-hop}}$ ,  $\mathbf{A}^{2\text{-hop}}$ ,  $\mathbf{A}^{\text{global}}$ ), and  $I = 8$  (per graph) for our Base model (and  $\hat{d} = 256$  for our Large model). Our FF sub-layer is a  $d$ -dimensional linear layer ( $d = 3\hat{d}$ ) followed by ReLU (Glorot et al., 2011) and dropout. The MLP Classifier consists of two  $4\hat{d}$ -dimensional linear layers with ReLU and dropout, followed by a 345-dimensional linear projection representing logits over the 345 categories in QuickDraw. We train all models by minimizing the softmax cross-entropy loss using the Adam (Kingma & Ba, 2014) optimizer for 100 epochs. We use an initial learning rate of  $5e-5$  and multiply by a factor 0.7 every 10 epochs. We use an early stopping strategy (with the hyper-parameter “patience” of 10 epochs) for selecting the final model, and the checkpoint with the highest validation performance is chosen to report test performance.

**Baselines** (i) From the perspective of coordinate-based sketch recognition, RNN models are a simple-yet-effective baseline. Following Xu et al. (2018), we design several bi-directional LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Cho et al., 2014) models at increasing parameter budgets comparable with MGT. The final RNN states are concatenated and passed to the MLP classifier described previously. We use batch size 256, initial learning rate  $1e-4$  and multiply by 0.9 every 10 epochs. We train models with both our multi-modal input (Section 3.2) as well as the 4D input from Xu et al. (2018).

(ii) Although converting sketch coordinates to images adds time overhead in practical settings and can be seen as auxiliary information, we compare MGT to various state-of-the-art CNN architectures. It is important to note that sketch sequences were truncated/padded for training both MGT and RNNs, hence image-based CNNs stand as an upper bound in terms of performance. We tune the optimal learning rate and decay schedule for each model, and use the maximum possible batch size. See details in Table 2. Following standard practice in computer vision (He et al., 2016; Huang et al., 2017), we employ early stopping based on observing over-fitting in the validation loss, and select the checkpoint with the highest validation accuracy for evaluation on the test set.

(iii) To evaluate the effectiveness of the proposed Graph Transformer layer, we compare it with popular GNN variants: the Graph Convolutional Network (Kipf & Welling, 2017) and the Graph Attention Network (Veličković et al., 2018)<sup>2</sup>. All GNN models follow the same hyperparameters and learning rate setup as Transformers ( $L = 4$ ,  $\hat{d} = 256$ ), and are augmented with residual connections and batch normalization for fair comparison, following (Bresson & Laurent, 2018).

<sup>2</sup>For GAT, we use the same scaled dot-product attention mechanism as GT as we found it to be significantly faster.

<sup>1</sup><https://quickdraw.withgoogle.com/data>

Table 2. Test set performance of MGT vs. the state-of-the-art RNN and CNN architectures. The  $1^{st}/2^{nd}/3^{rd}$  best results per column are indicated in red/blue/magenta.

Network	Configurations	Recognition Accuracy			Parameter Amount
		acc.@1	acc.@5	acc.@10	
Bi-directional LSTM #1	4D Input, $\hat{d} = 256, L = 4, drop_{RNN} = 0.5, drop_{MLP} = 0.15$	0.6665	0.8820	0.9189	5,553,241
Bi-directional LSTM #2	4D Input, $\hat{d} = 256, L = 5, drop_{RNN} = 0.5, drop_{MLP} = 0.15$	0.6524	0.8697	0.9133	7,130,201
Bi-directional GRU	4D Input, $\hat{d} = 256, L = 5, drop_{RNN} = 0.5, drop_{MLP} = 0.15$	0.6768	0.8854	0.9234	5,419,097
AlexNet (Krizhevsky et al., 2012)	LR:1e-4, dec. $\times$ 0.1 every 10 eps.	0.6808	0.8847	0.9203	58,417,305
VGG-11 (Simonyan & Zisserman, 2014)	LR:1e-2, dec. $\times$ 0.1 every 10 eps.	0.6743	0.8814	0.9191	130,179,801
Inception V3 (Szegedy et al., 2016)	LR:1e-3, dec. $\times$ 0.5 every 10 eps.	<b>0.7422</b>	<b>0.9189</b>	<b>0.9437</b>	25,315,474
ResNet-18 (He et al., 2016)	LR:1e-3, dec. $\times$ 0.7 every 5 eps.	0.7164	0.9072	0.9381	11,353,497
ResNet-34 (He et al., 2016)	LR:1e-3, dec. $\times$ 0.7 every 5 eps.	0.7154	0.9083	0.9375	21,461,657
ResNet-152 (He et al., 2016)	LR:1e-2, dec. $\times$ 0.1 every 10 eps.	0.6924	0.8973	0.9312	58,850,713
DenseNet-201 (Huang et al., 2017)	LR:1e-2, dec. $\times$ 0.1 every 10 eps.	0.7050	0.9013	0.9331	18,755,673
MobileNet V2 (Sandler et al., 2018)	LR:1e-3, dec. $\times$ 0.5 every 10 eps.	<b>0.7310</b>	<b>0.9161</b>	<b>0.9429</b>	2,665,817
Vanilla Transformer (Vaswani et al., 2017)	$\hat{d} = 256, L = 4, I = 8, drop = 0.1$ , Fully-connected graph	0.5249	0.7802	0.8486	14,029,401
MGT (Base)	$\hat{d} = 128, L = 4, I = 24, drop = 0.1, \mathbf{A}^{1-hop}, \mathbf{A}^{2-hop}, \mathbf{A}^{global}$	0.7070	0.9030	0.9351	10,096,601
MGT (Large)	$\hat{d} = 256, L = 4, I = 24, drop = 0.25, \mathbf{A}^{1-hop}, \mathbf{A}^{2-hop}, \mathbf{A}^{global}$	<b>0.7280</b>	<b>0.9106</b>	<b>0.9387</b>	39,984,729

## 4.2. Results

For fair comparison with RNN and CNN baselines at various parameter budgets, we implement two configurations of MGT: Base (10M parameters) and Large (40M parameters). Additionally, we perform several ablation studies to evaluate the effectiveness of our multi-graph architecture and our sketch-specific input design. Our main results are presented in Table 2.

**Comparison with RNN Baselines** We trained RNNs at various parameter budgets, and present results for the best performing bi-directional LSTM/GRU models in Table 2: (i) MGT outperforms both LSTM and GRU baselines by a significant margin (by 3% acc.@1 for Base, 5% for Large), indicating that both geometry and temporal order of strokes are important for sketch representation learning. (ii) Training larger RNNs is harder to converge, leading to degrading performance, *e.g.*, GRUs outperform deeper LSTMs by 2%. This result is not surprising: RNNs are notoriously hard to train at scale (Pascanu et al., 2013), while Transformer performance is known to improve with scale (Shoeybi et al., 2019).

**Comparison with CNN Baselines** Table 2 also presents performance of several state-of-the-art CNN architectures for computer vision:

(i) Inception V3 (Szegedy et al., 2016) and MobileNet V2 (Sandler et al., 2018) are the best performing CNN architectures. Our MGT Base has competitive recognition accuracy to all other baselines: AlexNet (Krizhevsky et al., 2012), VGG-11 (Simonyan & Zisserman, 2014), the ResNet family (He et al., 2016), and DenseNet-201 (Huang et al., 2017). (ii) MGT Large has small performance gap to Inception V3 and MobileNet V2 (*i.e.*, 72.80% acc.@1 vs. 74.22%,

72.80% acc.@1 vs. 73.10%) and outperforms all other CNN architectures by 1%.

(iii) Somewhat counter-intuitively, shallow networks (Inception V3, MobileNet V2) outperform deeper networks (ResNet-152, Densenet-201). This result highlights that CNNs designed for images with dense colors and textures are un-suitable for sparse sketches.

Note that MobileNet V2 is specifically designed for fast inference on mobile phones and is not directly comparable in terms of model parameters.

**Ablations for Multi-Graph Architecture** We design several ablation studies to evaluate our sketch-specific multi-graph architecture in Table 3:

(i) We evaluate Graph Transformers trained on fully-connected graphs, *i.e.* vanilla Transformers (GT #1), fully-connected graphs within strokes (GT #2), as well as random graphs with 10%, 20% and 30% connectivity (GT #3, #4, and #5 respectively). We compare their performance with Graph Transformers trained on sketch-specific graphs  $\mathbf{A}^{1-hop}$  (GT #6),  $\mathbf{A}^{2-hop}$  (GT #7),  $\mathbf{A}^{3-hop}$  (GT #8), and  $\mathbf{A}^{global}$  (GT #9). We find that vanilla Transformers on fully-connected (52.49% acc.@1) and random graphs (52.71%, 53.52%, 53.22%) perform poorly compared to sketch-specific graph structures determined by domain expertise, such as fully-connected stroke graphs (64.87%) and  $\mathbf{A}^{1-hop}$  (70.23%). The superior performance of  $K$ -hop graphs suggests that Transformers benefit from sparse graphs representing local sketch geometry. We also evaluate a combined sketch-specific graph structure, *i.e.*,  $\mathbf{A}^{1-hop} \cup \mathbf{A}^{2-hop} \cup \mathbf{A}^{global}$  (GT #10), where the graph connectivity is the logical union set of  $\mathbf{A}^{1-hop}$ ,  $\mathbf{A}^{2-hop}$ , and  $\mathbf{A}^{global}$ . However, this structure fails to gain performance improvement over  $\mathbf{A}^{1-hop}$  and  $\mathbf{A}^{2-hop}$ , despite involving more domain knowledge.

(ii) Next, we experiment with various permutations of

Table 3. Ablation study for multi-graph architecture of MGT. GT denotes single-graph variants of MGT. The 1<sup>st</sup>/2<sup>nd</sup> best results per column are indicated in red/blue.  $\cup$  denotes the logical union operation.

Network	Configurations						Recognition Accuracy			Parameter Amount
	$G$	Graph Structure	$I_{total}$	$\tilde{d}$	$L$	dropout	acc.@1	acc.@5	acc.@10	
GT #1	1	Fully conn. ( <i>vanilla</i> )	8	256	4	0.10	0.5249	0.7802	0.8486	14,029,401
GT #2	1	Intra-stroke fully conn.	8	256	4	0.10	0.6487	0.8697	0.9151	14,029,401
GT #3	1	Random (10%)	8	256	4	0.10	0.5271	0.7890	0.8589	14,029,401
GT #4	1	Random (20%)	8	256	4	0.10	0.5352	0.7945	0.8617	14,029,401
GT #5	1	Random (30%)	8	256	4	0.10	0.5322	0.7917	0.8588	14,029,401
GT #6	1	$\mathbf{A}^{1-hop}$	8	256	4	0.10	0.7023	0.8974	0.9303	14,029,401
GT #7	1	$\mathbf{A}^{2-hop}$	8	256	4	0.10	0.7082	0.8999	0.9336	14,029,401
GT #8	1	$\mathbf{A}^{3-hop}$	8	256	4	0.10	0.7028	0.8991	0.9327	14,029,401
GT #9	1	$\mathbf{A}^{global}$	8	256	4	0.10	0.5488	0.8009	0.8659	14,029,401
GT #10	1	$\mathbf{A}^{1-hop} \cup \mathbf{A}^{2-hop} \cup \mathbf{A}^{global}$	8	256	4	0.10	0.7057	0.9021	0.9346	14,029,401
MGT #11	2	$\mathbf{A}^{1-hop}, \mathbf{A}^{2-hop}$	16	256	4	0.25	0.7149	0.9049	0.9361	28,188,249
MGT #12	2	$\mathbf{A}^{1-hop}, \mathbf{A}^{global}$	16	256	4	0.25	0.7111	0.9041	0.9355	28,188,249
MGT #13	2	$\mathbf{A}^{2-hop}, \mathbf{A}^{global}$	16	256	4	0.25	<b>0.7237</b>	<b>0.9102</b>	<b>0.9400</b>	28,188,249
MGT #14	3	$\mathbf{A}^{1-hop}, \mathbf{A}^{1-hop}, \mathbf{A}^{1-hop}$	24	256	4	0.25	0.7077	0.9020	0.9340	39,984,729
MGT #15	3	$\mathbf{A}^{1-hop}, \mathbf{A}^{2-hop}, \mathbf{A}^{3-hop}$	24	256	4	0.25	0.7156	0.9066	0.9365	39,984,729
MGT #16	3	$\mathbf{A}^{1-hop} \cup \mathbf{A}^{2-hop} \cup \mathbf{A}^{global}$	24	256	4	0.25	0.7126	0.9051	0.9372	39,984,729
MGT #17	3	$\mathbf{A}^{1-hop}, \mathbf{A}^{2-hop}, \mathbf{A}^{global}$	24	256	4	0.25	<b>0.7280</b>	<b>0.9106</b>	<b>0.9387</b>	39,984,729

graphs for multi-graph models (MGT #11-#17). We find that using a 3-graph architecture (MGT #17) combining local sketch geometry ( $\mathbf{A}^{1-hop}, \mathbf{A}^{2-hop}$ ) and global temporal relationships ( $\mathbf{A}^{global}$ ) significantly boosts performance over 2-graph and 1-graph models (72.80% vs. 72.37% for 2-graph and 70.82% for 1-graph). This result is interesting because using global graphs independently (GT #9) leads to comparatively poor performance (54.88%). Additionally, we found that using diverse graphs (MGT #15, #17) is better than using the same graph (MGT #14). Comparing MGT #14 and MGT #6 further shows that performance gains are due to the multi-graph architecture as opposed to more model parameters.

(iii) We also repeatedly input the adjacency matrix of GT #10 (*i.e.*,  $\mathbf{A}^{1-hop} \cup \mathbf{A}^{2-hop} \cup \mathbf{A}^{global}$ ) three times as the multiple graph structures to train our MGT (see MGT #16 in Table 3). Compared with MGT #17, there is a clear performance gap (71.26% vs. 72.80%). This further validates our idea of learning sketch representations through multiple separate graphs.

**Comparison with GNN Baselines** In Table 4, we present performance of our Graph Transformer layer compared to GCN and GAT, two popular GNN variants:

- (i) We find that all models perform similarly on fully-connected graphs. Using 1-hop graphs results in significant gains for all models, with Transformer performing the best.
- (ii) Interestingly, both GNNs on fully-connected graphs are outperformed by a simple position-wise embedding method without any graph structure: each node undergoes 4 feed-forward (FF) layers followed by summation and the MLP classifier. These results further highlight the importance of

Table 4. Test set performance of Graph Transformer vs. other GNN variants. The 1<sup>st</sup>/2<sup>nd</sup> best results per column are indicated in red/blue. ‘‘Gra. Stru.’’ denotes graph structure.

Network	Gra. Stru.	Recognition Accuracy			Parameter Amount
		acc.@1	acc.@5	acc.@10	
GCN	Full	0.4098	0.7384	0.8213	6,948,441
	$\mathbf{A}^{1-hop}$	0.6800	0.8869	0.9224	
GAT	Full	0.4098	0.6960	0.7897	11,660,889
	$\mathbf{A}^{1-hop}$	<b>0.6977</b>	<b>0.8952</b>	<b>0.9298</b>	
GT	Full	0.5242	0.7796	0.8465	14,029,401
	$\mathbf{A}^{1-hop}$	<b>0.7057</b>	<b>0.8992</b>	<b>0.9311</b>	
Ps.-wise FF	None	0.5296	0.7901	0.8576	4,586,073

sketch-specific graph structures for the success of Transformers. Our final models use the Transformer layer, which implicitly includes the FF sub-layer, Eq. (7).

**Ablations for Multi-Modal Input** In Table 5, we experiment with various permutations of our sketch-specific multi-modal input design. We aggregate information from spatial (coordinates), semantic (flag bits), and temporal (position encodings) modalities via summation (as in Transformers for NLP) or concatenation:

- (i) Effectively using all modalities is important for performance (*e.g.*, ‘‘C(coo., flag, pos.)’’ outperforms ‘‘coo.’’ and ‘‘C(coo., flag)’’: 72.80% acc.@1 vs. 65.12%, 70.17%).
- (ii) Concatenation works better than 4D input as well as summation (*e.g.*, ‘‘C(coo., flag, pos.)’’ outperforms ‘‘C(4D Input, pos.)’’ and ‘‘coo. + flag + pos.’’: 72.80% acc.@1 vs.

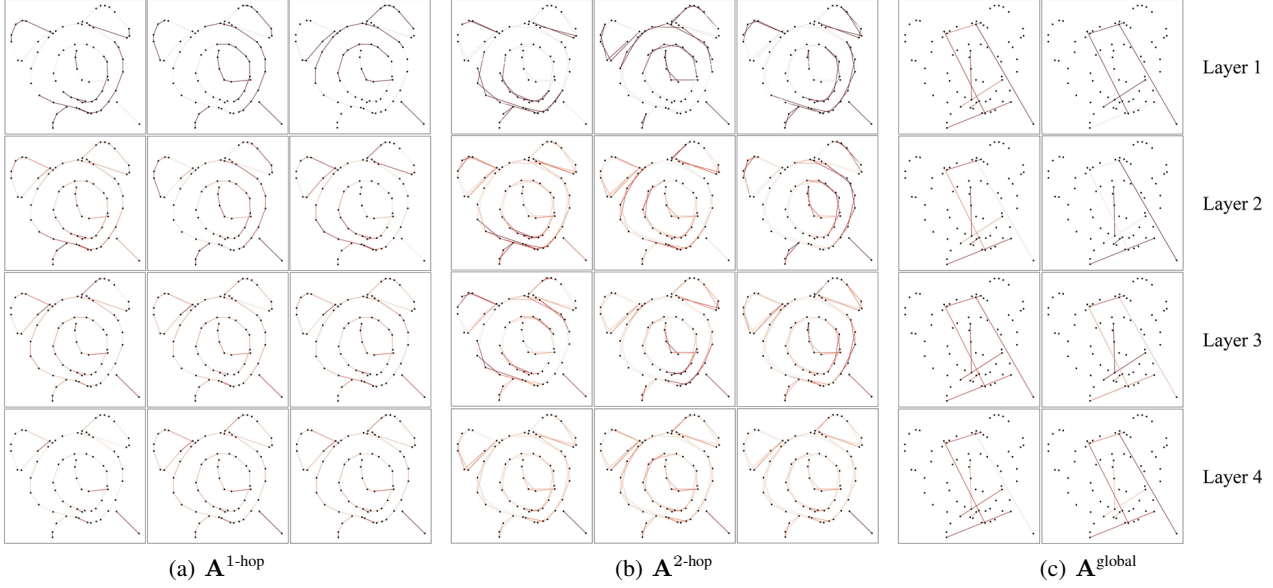


Figure 6. Selected attention heads at each layer of MGT for a sample from the test set (labelled ‘alarm clock’). Each layer has  $I = 8$  attention heads per graph in total. We manually choose the most interesting heads for each graph. Darker reds indicate higher attention values. Best viewed in color.

Table 5. Ablation study for multi-modal input for MGT (Large). Notations: “+” and “ $\mathcal{C}(\dots)$ ” denote “sum” and “concatenate”, respectively; “coo.”, “flag”, and “pos.” represent “coordinate”, “flag bit”, and “position encoding”, respectively. The 1<sup>st</sup>/2<sup>nd</sup> best results per column are indicated in red/blue.

Input Permutation	Recognition Accuracy		
	acc.@1	acc.@5	acc.@10
coo.	0.6512	0.8735	0.9162
coo. + flag	0.6568	0.8762	0.9176
coo. + flag + pos.	0.6600	0.8766	0.9182
$\mathcal{C}(\text{coo.}, \text{flag})$	0.7017	0.8996	0.9321
$\mathcal{C}(\text{coo.}, \text{flag}, \text{pos.})$	<b>0.7280</b>	<b>0.9106</b>	<b>0.9387</b>
4D Input	0.6559	0.8758	0.9175
4D Input + pos.	0.6606	0.8781	0.9190
$\mathcal{C}(4\text{D Input}, \text{pos.})$	<b>0.7117</b>	<b>0.9048</b>	<b>0.9366</b>

71.17%, 66.06%).

**Qualitative Results** In Figure 6, we visualize attention heads at each layer of MGT for a sample from the test set (labelled ‘alarm clock’). Attention heads in the initial layers attend very strongly to certain neighbors and very weakly to others, *i.e.*, the model builds local patterns for sketch sub-components (strokes) through message passing along their contours. In penultimate layers, the intensity of neighborhood attention is significantly lower and evenly distributed, indicating that the model is aggregating information from various strokes at each node. We believe  $A^{\text{global}}$  graphs aid in message passing between strokes, enabling the model to

understand their relative arrangement, *e.g.*, the feet of the clock are attached to the bottom of the body, the arms are located inside the body, *etc.*

Additional visualization results are available in the Supplementary Material along with preliminary experiments for tuning baseline models and MGT.

## 5. Conclusion

This paper introduces a novel representation of free-hand sketches as multiple sparsely connected graphs. We design a Multi-Graph Transformer (MGT) for capturing both geometric structure and temporal information from sketch graphs. The intrinsic traits of the MGT architecture include: (i) using graphs as universal representations of sketch geometry, as well as temporal and semantic information, (ii) injecting domain knowledge into Transformers through sketch-specific graphs, and (iii) making full use of multiple intra-stroke and extra-stroke graphs. In future work, we shall explore more nuanced graph structures for free-hand sketches, in order to further inject domain knowledge into the MGT architecture.

We hope MGT can serve as a foundation for future work in sketch applications and network architectures, motivating the community towards sketch representation learning using graphs. Additionally, for the graph neural network (GNN) community, we hope that MGT helps free-hand sketch become a new test-bed for GNNs.



## Acknowledgements

Xavier Bresson is supported in part by NRF Fellowship NRFF2017-10.

## References

- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- Bresson, X. and Laurent, T. An experimental study of neural networks for variable graphs. In *ICLR Workshop*, 2018.
- Bruna, J., Zaremba, W., Szlam, A., and Lecun, Y. Spectral networks and locally connected networks on graphs. In *ICLR*, 2014.
- Chen, W. and Hays, J. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *CVPR*, 2018.
- Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
- Collomosse, J., Bui, T., and Jin, H. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019.
- Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.
- Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, 2016.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.
- Dey, S., Riba, P., Dutta, A., Lladós, J., and Song, Y.-Z. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019.
- Dutta, A. and Akata, Z. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019.
- Edunov, S., Ott, M., Auli, M., and Grangier, D. Understanding back-translation at scale. In *EMNLP*, 2018.
- Eitz, M., Hays, J., and Alexa, M. How do humans sketch objects? *ACM TOG*, 2012.
- Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *AISTATS*, 2011.
- Ha, D. and Eck, D. A neural representation of sketch drawings. In *ICLR*, 2018.
- Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *NeurIPS*, 2017.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 1997.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.
- Liu, F., Deng, X., Lai, Y.-K., Liu, Y.-J., Ma, C., and Wang, H. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In *CVPR*, 2019.
- Liu, L., Shen, F., Shen, Y., Liu, X., and Shao, L. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017.
- Lu, Y., Wu, S., Tai, Y.-W., and Tang, C.-K. Image generation from sketch constraint using contextual gan. In *ECCV*, 2018.
- Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*, 2017.
- Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *ICML*, 2013.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018.

- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.
- Sangkloy, P., Burnell, N., Ham, C., and Hays, J. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016.
- Sarvadevabhatla, R. K., Kundu, J., and Babu R, V. Enabling my robot to play pictonary: Recurrent neural networks for sketch recognition. In *ACM MM*, 2016.
- Shen, Y., Liu, L., Shen, F., and Shao, L. Zero-shot sketch-image hashing. In *CVPR*, 2018.
- Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Song, J., Yu, Q., Song, Y.-Z., Xiang, T., and Hospedales, T. M. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014.
- Sukhbaatar, S., Szlam, A., and Fergus, R. Learning multi-agent communication with backpropagation. In *NeurIPS*, 2016.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph Attention Networks. In *ICLR*, 2018.
- Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., and Chao, L. S. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.
- Xu, P., Huang, Y., Yuan, T., Pang, K., Song, Y.-Z., Xiang, T., Hospedales, T. M., Ma, Z., and Guo, J. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *CVPR*, 2018.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.
- Ye, Y., Lu, Y., and Jiang, H. Human’s scene sketch understanding. In *ICMR*, 2016.
- Ye, Z., Guo, Q., Gan, Q., Qiu, X., and Zhang, Z. Bp-transformer: Modelling long-range context via binary partitioning. *arXiv preprint arXiv:1911.04070*, 2019.
- Yu, Q., Yang, Y., Song, Y.-Z., Xiang, T., and Hospedales, T. Sketch-a-net that beats humans. In *BMVC*, 2015.

## A. Impact of Scaling RNNs and Transformers

To explore the impact of model scale on coordinate-based architectures, we explored training Bi-directional GRUs and GTs at various parameter budgets, see Table 6. We used  $A^{1\text{-hop}}$  graphs for GTs for fair comparison:  $A^{1\text{-hop}}$  graphs are the most semantically similar to the bi-directionally sequential processing of input coordinates in the GRU models.

Table 6. Validation set performance of GT and RNN architectures at various parameter budgets. Best results for each model are indicated in **bold**.

Network	Configurations	acc.@1	Parameters
Bi-direct GRU	Trf. Input, $\hat{d} = 64, L = 2, \text{drop} = 0.1$	0.6100	1,370,265
	Trf. Input, $\hat{d} = 64, L = 4, \text{drop} = 0.1$	<b>0.6592</b>	2,701,977
	Trf. Input, $\hat{d} = 128, L = 2, \text{drop} = 0.1$	0.6108	5,279,705
	Trf. Input, $\hat{d} = 128, L = 4, \text{drop} = 0.1$	0.6436	10,597,337
	Trf. Input, $\hat{d} = 256, L = 2, \text{drop} = 0.1$	0.5104	20,717,145
	Trf. Input, $\hat{d} = 256, L = 4, \text{drop} = 0.1$	0.6105	41,969,241
GT $A^{1\text{-hop}}$ graph	$\hat{d} = 64, L = 2, I = 4, \text{drop} = 0.1$	0.5332	581,145
	$\hat{d} = 64, L = 4, I = 4, \text{drop} = 0.1$	0.5872	951,705
	$\hat{d} = 128, L = 2, I = 4, \text{drop} = 0.1$	0.6334	2,128,601
	$\hat{d} = 128, L = 4, I = 4, \text{drop} = 0.1$	0.6626	3,607,001
	$\hat{d} = 128, L = 6, I = 4, \text{drop} = 0.1$	0.6759	5,085,401
	$\hat{d} = 128, L = 8, I = 4, \text{drop} = 0.1$	0.6857	6,563,801
	$\hat{d} = 256, L = 2, I = 8, \text{drop} = 0.1$	0.6835	8,123,481
	$\hat{d} = 256, L = 4, I = 8, \text{drop} = 0.1$	0.7057	14,029,401
	$\hat{d} = 256, L = 6, I = 8, \text{drop} = 0.1$	<b>0.7163</b>	19,935,321
	$\hat{d} = 512, L = 4, I = 16, \text{drop} = 0.1$	0.7025	55,321,433

At low parameter budgets, GRUs outperform to GTs. Interestingly, GRU performance degrades as parameter counts rise. On the other hand, GT performance consistently improved with scale, plateauing at approximately 70% validation accuracy. Our results correspond with studies on scaling RNNs and Transformers for NLP (Pascanu et al., 2013; Shoeybi et al., 2019). In future work, it will be interesting to explore the impact of scaling the dataset instead, i.e., using all 50 million sketches from Google QuickDraw.

## B. Using Generic Sparse Graphs

To further emphasize the impact of sketch-specific graph structures for the successful application of GNN-based models, we compare the use of *generic* sparse graph structures with sketch-specific graphs in Table 7. We found using fully connected and random sparse graphs lead to poor performance. Using the Euclidean  $k$ -nearest neighbors for graph sparsification boosted accuracy by approximately 5%. Incorporating sketch-specific domain knowledge into graph construction, *e.g.*,  $\mathbf{A}^{k\text{-hop}}$  graphs, substantially improved performance over generic sparsification techniques.

Table 7. Validation set performance of GT model ( $\hat{d} = 256$ ,  $L = 4$ ,  $I = 8$ ,  $drop = 0.1$ ) using random, generic and sketch-specific graph structures.

Graph Structure	acc.@1
Fully connected	0.5249
Random (10%)	0.5271
Random (20%)	0.5352
Euclidean 5-nn.	0.5934
Euclidean 10-nn.	0.5877
Euclidean 25-nn.	0.5649
$\mathbf{A}^{1\text{-hop}}$	0.7023
$\mathbf{A}^{2\text{-hop}}$	0.7082
$\mathbf{A}^{3\text{-hop}}$	0.7028

Overall, our results suggest that merely sparse graphs are not sufficient for the application of GNNs to sketches. We advocate for the injection of domain knowledge into GNNs through the use of sketch-specific sparse graphs.

## C. Tuning Learning Rates

We found tuning the learning rate schedule to be essential for the performance of all models. Table 8 presents the validation set performance for GT and CNN architectures for various learning rate schedules. As one would expect, image-based CNNs are able to learn at very high learning rates, fitting the training data in 10-20 epochs. On the other hand, coordinate-based RNN and GT architectures require lower initial learning rates for convergence, thus taking more epochs. Our final RNN models use an initial learning rate  $1e-4$  and decay by 0.9 every 10 epochs, while our final GT models use an initial learning rate  $5e-5$  and multiply by 0.7 every 10 epochs. For CNNs, the optimal learning rate schedule is chosen based on validation set performance for each architecture.

Future work shall explore the use of learning rate warmup strategies, which have been effectively used to train large-scale Transformer models (Vaswani et al., 2017; Radford et al., 2018).

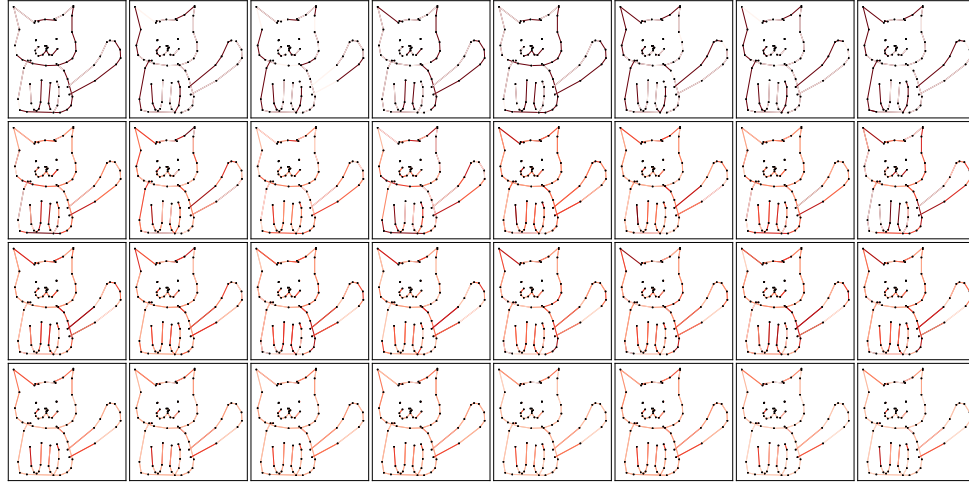
Table 8. Validation set performance of GT and CNN architectures for various learning rate schedules. Best results for each model are indicated in **bold**.

Network	Configurations	acc.@1	Parameters
AlexNet	LR: $1e-4$ , dec. $\times 0.1$ every 10 eps.	0.6808	58,417,305
	LR: $1e-4$ , dec. $\times 0.5$ every 10 eps	<b>0.6750</b>	
	LR: $1e-4$ , dec. $\times 0.7$ every 5 eps	0.6651	
	LR: $1e-4$ , dec. $\times 0.9$ every 1 eps	0.6630	
	LR: $5e-5$ , dec. $\times 0.9$ every 1 eps	0.6480	
Inception V3	LR: $1e-2$ , dec. $\times 0.1$ every 10 eps	0.7262	25,315,474
	LR: $1e-2$ , dec. $\times 0.9$ every 1 eps.	0.7227	
	LR: $1e-2$ , dec. $\times 0.1$ every 1 eps.	0.6213	
	LR: $1e-3$ , dec. $\times 0.5$ every 10 eps.	<b>0.7426</b>	
	LR: $1e-3$ , dec. $\times 0.7$ every 5 eps.	0.7335	
ResNet-18	LR: $1e-2$ , dec. $\times 0.5$ every 10 eps	0.6907	11,353,497
	LR: $1e-2$ , dec. $\times 0.1$ every 10 eps	0.7031	
	LR: $1e-2$ , dec. $\times 0.9$ every 1 eps	0.6977	
	LR: $1e-2$ , dec. $\times 0.1$ every 1 eps	0.6768	
	LR: $1e-3$ , dec. $\times 0.5$ every 10 eps	0.7036	
	LR: $1e-3$ , dec. $\times 0.7$ every 5 eps	<b>0.7164</b>	
	LR: $1e-4$ , dec. $\times 0.9$ every 10 eps	0.7006	
ResNet-34	LR: $1e-2$ , dec. $\times 0.1$ every 10 eps	0.7009	21,461,657
	LR: $1e-2$ , dec. $\times 0.9$ every 1 eps	0.7018	
	LR: $1e-2$ , dec. $\times 0.1$ every 1 eps	0.6773	
	LR: $1e-3$ , dec. $\times 0.5$ every 10 eps	0.7116	
	LR: $1e-3$ , dec. $\times 0.7$ every 5 eps	<b>0.7154</b>	
MobileNet V2	LR: $1e-2$ , dec. $\times 0.1$ every 10 eps	0.7106	2,665,817
	LR: $1e-2$ , dec. $\times 0.5$ every 10 eps	0.7241	
	LR: $1e-2$ , dec. $\times 0.9$ every 10 eps	0.7236	
	LR: $1e-3$ , dec. $\times 0.5$ every 10 eps	<b>0.7305</b>	
	LR: $1e-3$ , dec. $\times 0.7$ every 5 eps	0.7237	
	LR: $1e-3$ , dec. $\times 0.9$ every 10 eps	0.7227	
	LR: $1e-3$ , dec. $\times 0.9$ every 1 eps	0.7204	
	LR: $1e-4$ , dec. $\times 0.9$ every 10 eps	0.6813	
	LR: $1e-3$ , dec. $\times 0.9$ every 1 eps	0.4032	
GT, $\hat{d} = 256$ , $L = 4$ , $I = 8$ , $drop = 0.1$ , $\mathbf{A}^{2\text{-hop}}$ graph	LR: $5e-4$ , dec. $\times 0.9$ every 1 eps	0.4522	14,029,401
	LR: $5e-4$ , dec. $\times 0.7$ every 5 eps	0.4546	
	LR: $5e-4$ , dec. $\times 0.5$ every 10 eps	0.5366	
	LR: $1e-4$ , dec. $\times 0.9$ every 1 eps	0.7050	
	LR: $1e-4$ , dec. $\times 0.7$ every 5 eps	0.7101	
	LR: $1e-4$ , dec. $\times 0.5$ every 10 eps	0.7101	
	LR: $5e-5$ , dec. $\times 0.9$ every 10 eps	0.7041	
	LR: $5e-5$ , dec. $\times 0.9$ every 5 eps	0.7105	
	LR: $5e-5$ , dec. $\times 0.9$ every 1 eps	0.6772	
	LR: $5e-5$ , dec. $\times 0.7$ every 10 eps	<b>0.7120</b>	

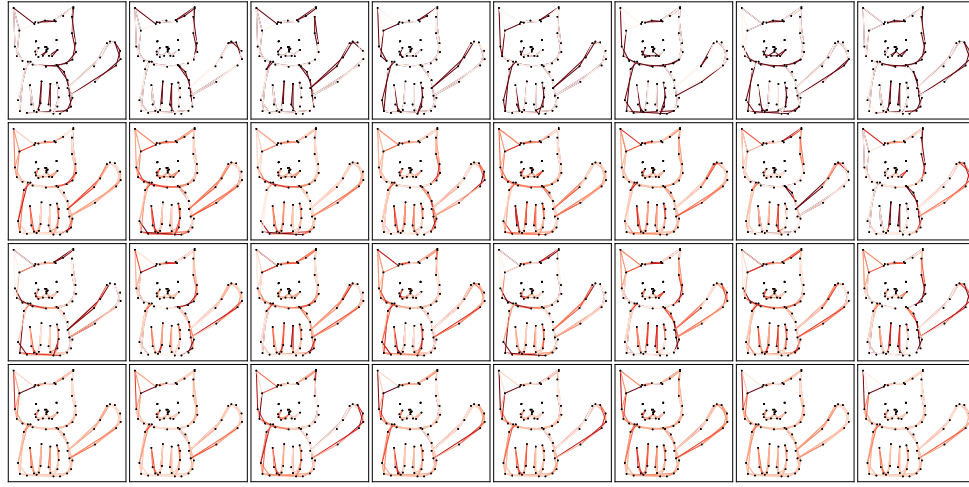
## D. Additional Attention Visualizations

In Figures 7, 8 and 9, we visualize attention heads at each layer of MGT for various test set samples. Each sub-figure contains attention heads for each of the three graphs ( $\mathbf{A}^{1\text{-hop}}$ ,  $\mathbf{A}^{2\text{-hop}}$ ,  $\mathbf{A}^{global}$ ), and each of the rows 1-4 in each sub-figure correspond to layers 1-4. Darker reds indicate higher attention values. All figures are best viewed in color.

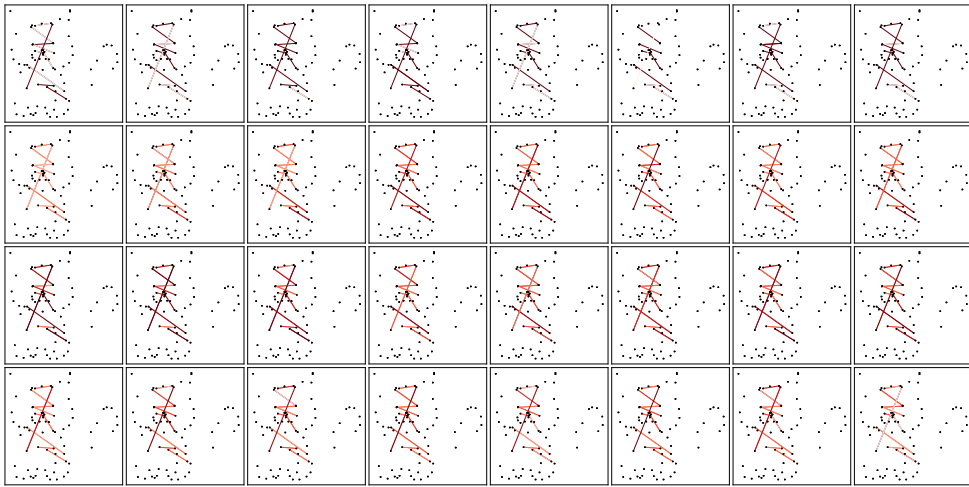
Attention heads in the initial layers attend very strongly to certain neighbors and very weakly to others, *i.e.*, the model builds local patterns for sketch sub-components (strokes) through message passing along their contours. In penultimate layers, the intensity of neighborhood attention is significantly lower and evenly distributed, indicating that the model is aggregating information from various strokes at each node.



(a)  $\mathbf{A}^{1\text{-hop}}$



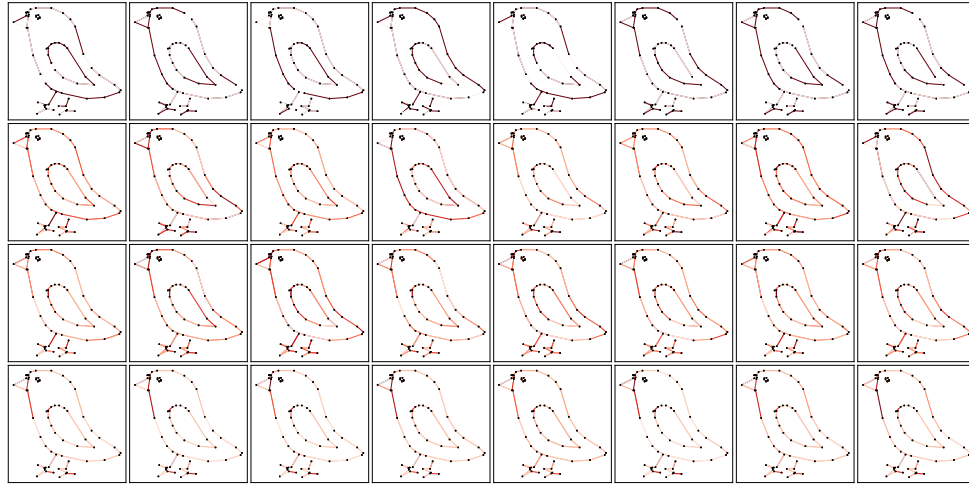
(b)  $\mathbf{A}^{2\text{-hop}}$



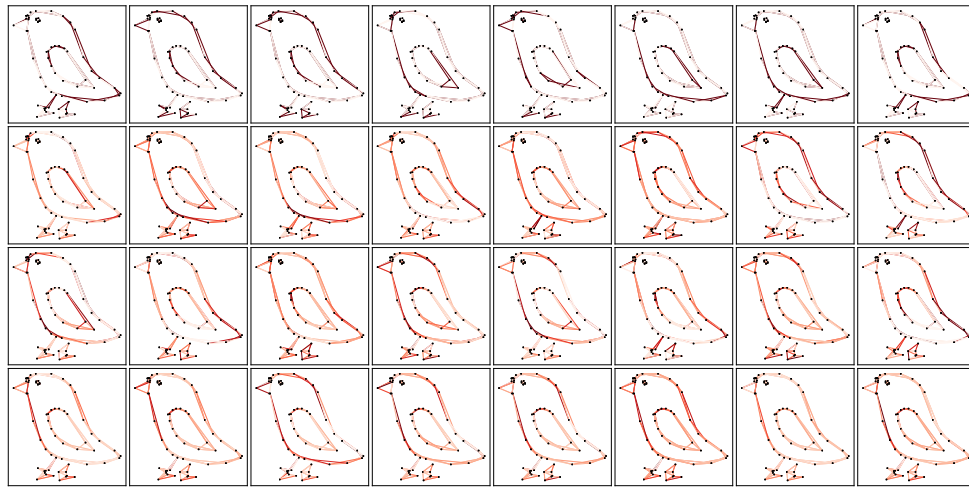
(c)  $\mathbf{A}^{global}$

Figure 7. Attention heads at each layer of MGT for a test set sample labelled *cat*.

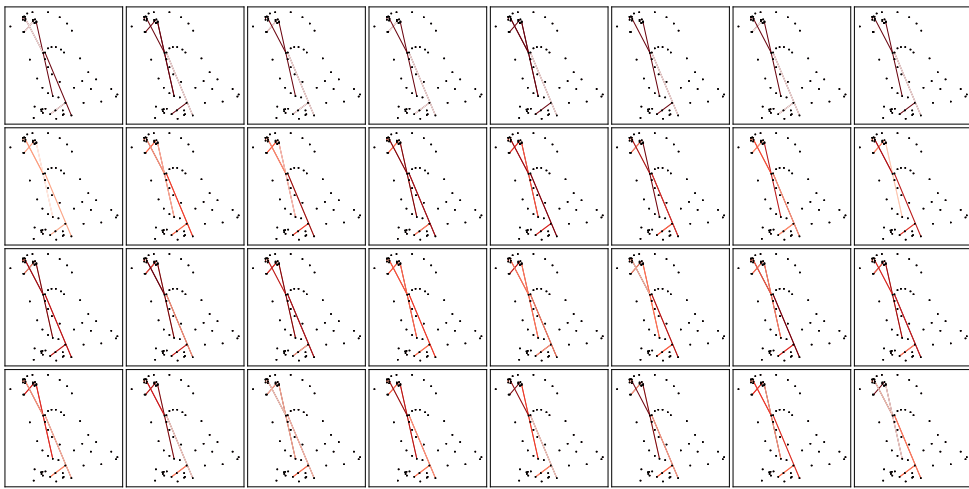




(a)  $A^{1\text{-hop}}$

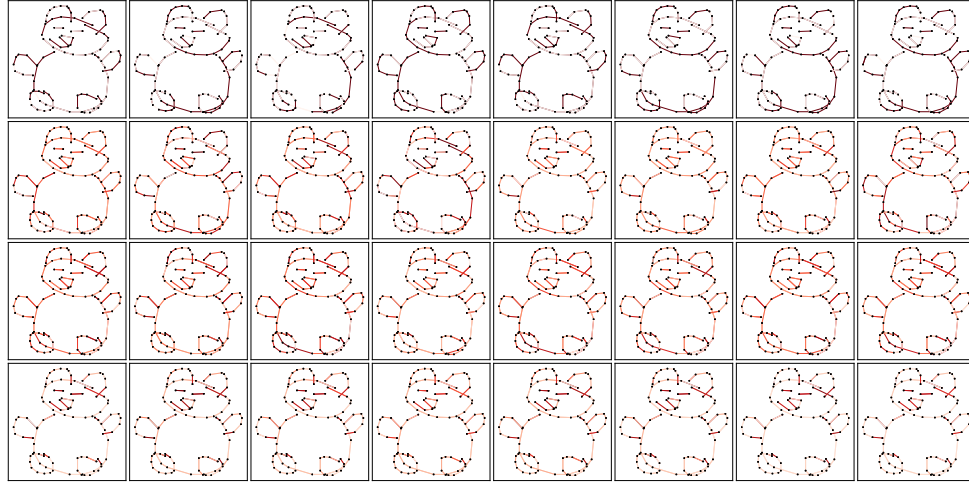


(b)  $A^{2\text{-hop}}$

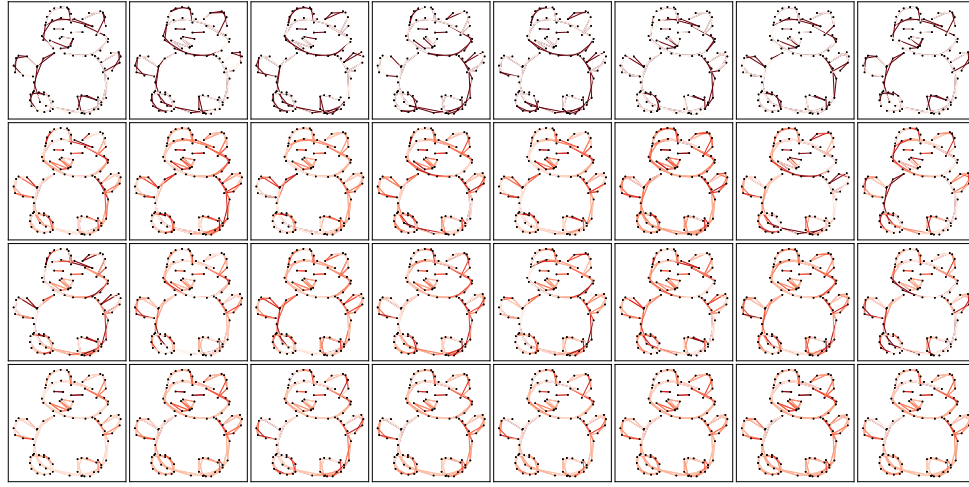


(c)  $A^{global}$

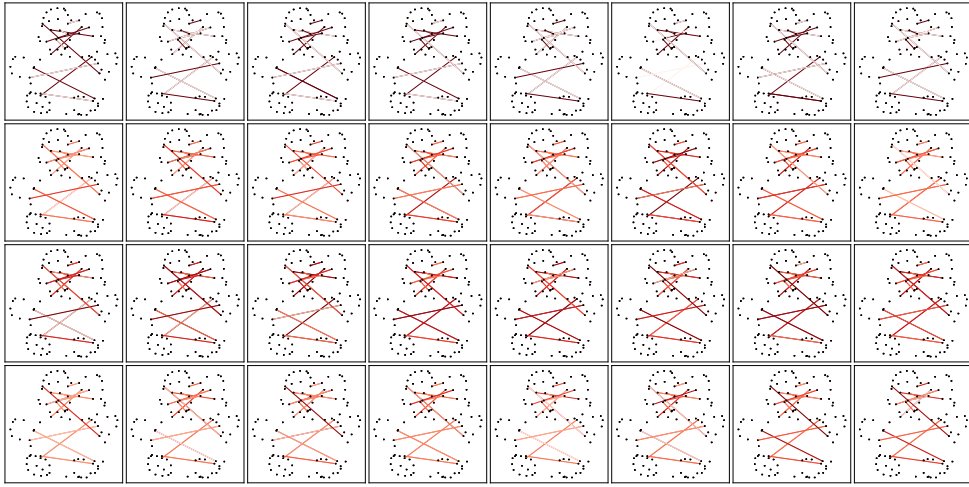
Figure 8. Attention heads at each layer of MGT for a test set sample labelled *bird*.



(a)  $\mathbf{A}^{1\text{-hop}}$



(b)  $\mathbf{A}^{2\text{-hop}}$



(c)  $\mathbf{A}^{global}$

Figure 9. Attention heads at each layer of MGT for a test set sample labelled *teddy*.