

# Graph autoencoder with constant dimensional latent space

Adam Małkowski, Jakub Grzechociński, Paweł Wawrzyński

**Abstract**—Invertible transformation of large graphs into constant dimensional vectors (embeddings) remains a challenge. In this paper we address it with recursive neural networks: The encoder and the decoder. The encoder network transforms embeddings of subgraphs into embeddings of larger subgraphs, and eventually into the embedding of the input graph. The decoder does the opposite. The dimension of the embeddings is constant regardless of the size of the (sub)graphs. Simulation experiments presented in this paper confirm that our proposed graph autoencoder can handle graphs with even thousands of vertices.

**Index Terms**—Graph Neural Network, Autoencoder.

## I. INTRODUCTION

Graph Neural Networks (Graph NNs, GNNs) (Wu et al., 2019; Zhou et al., 2019) is an emerging area within artificial intelligence. It addresses operations on graphs such as their generation, representation, classification, as well as operations on their separate nodes or edges such as classification or prediction of their attributes.

In this paper, we design a transformation (encoding) of a set of graphs into vectors of fixed size (embeddings) and inverse transformation (decoding). Our proposed transformations may be applied, among others, to (i) graph classification, with the fixed-size graph embedding fed to an ordinary classifier, (ii) graph evaluation/labeling, with the fixed-size graph embedding fed to a general purpose function approximator or (iii) graph generation, with a noise vector fed to the decoder.

Both proposed transformations are based on feedforward NNs applied recursively to embeddings of subgraphs of the input graph. The encoder recursively aggregates embeddings of subgraphs into embeddings of larger subgraphs. The decoder, conversely, recursively desegregates the embeddings and produces the elements of the adjacency matrix.

In the literature, graphs are typically represented with arrays of embeddings of their vertices. These structures are impossible to handle with methods that accept input of fixed size, such as feedforward neural networks, as they have a different shape depending on the graph size. There are also methods that embed graphs in vectors of fixed size, but they do not enable reconstruction of the graphs from these vectors, hence they lose some information on these graphs.

To the best of our knowledge, our method is the first one able to represent graphs of arbitrary sizes with embeddings of fixed dimension which enables reconstruction of the source graphs thereby preserving all the information on these graphs.

The paper is organized as follows. Sec. II overviews related literature. Sec. III introduces our solution. Sec. IV presents an experimental study, and Sec. V concludes the paper.

*a) Formal problem description.*: We consider undirected graphs. A graph of interest is given by a number of its vertices,  $n \in \mathbb{N}$ , and its adjacency matrix  $A \in \{0, 1\}^{n \times n}$ . Its entry  $A_{i,j}$  indicates if there is an edge between the vertices  $i$  and  $j$ .

For a given set of graphs and  $m \in \mathbb{N}$  we seek for a transformation of each graph in this set into a vector in  $\mathbb{R}^m$ , and the inverse transformation.

Our proposed solution in its extended form presented in Sec. III-C applies also to more general problems with directed graphs, weighted edges, and labeled edges and vertices.

## II. RELATED WORK

*a) Graph embeddings in vectors of fixed size.*: These methods embed graphs in vectors in  $\mathbb{R}^m$  for a fixed  $m$ . Their primary goal is to represent graphs in a set of features to enable their classification. Methods like Graph Kernels (Yanardag & Vishwanathan, 2015a) Graph2Vec (Narayanan et al., 2020) are inspired by natural language processing techniques and describe graphs with a concentration of specific subgraphs in there. Later methods define NNs that convert graphs into embeddings in  $\mathbb{R}^m$  with a neural network learning directly to optimize the graph classification criterion. In DiffPool (Ying et al., 2018) the NN operates on a hierarchy of subgraphs. CapsGNN (Zhang & Chen, 2019) is a combination of graph NN and a capsule NN. UGraphEmb (Bai et al., 2019) use multiscale node attention and graph proximity metrics to assure that a distance between graphs corresponds to the distance between their embeddings.

*b) Graph representation with vertex embeddings.*: In the literature on graph representation, an embedding of a graph is a collection of embeddings of its vertices. They have common size,  $d \in \mathbb{N}$ . A graph embedding results from training on this single graph. The main purpose of this training is to enable classification of vertices. All papers discussed in this paragraph (except Li et al., 2020) use the term “graph embedding” this way.

Cao et al. (2016) introduced a method that designed graph embeddings based on a random walk through the graph. Wang et al. (2016) considered graphs with weighted edges. They defined embeddings of vertices that reflected 1-st and 2-nd proximity between the vertices. Yang et al. (2016) proposed embeddings of vertices optimized to reflect the context of vertices as well as their labels.

Hamilton et al. (2017) introduced GraphSAGE. This method embeds vertices with vectors that are a function of the embeddings of their neighbors. This concept raised an issue if vertex embeddings coupled this way remain in a steady state. This issue was addressed by Dai et al. (2018). Klicpera

et al. (2019) proposed vertex embeddings being a result of steady-state message passing between neighboring and more distant vertices.

Chami et al. (2019) introduced hyperbolic geometry instead of Euclidean one to embed vertices in the graph. That improved the performance of the resulting graph convolutional NN. (Mehta et al., 2019) proposed vertex embeddings resulting from a combination of graph convolution networks and stochastic blockmodels introduced earlier for latent space modeling (Nowicki & Snijders, 2001).

Velickovic et al. (2019) proposed vertex embeddings that summarized the neighborhood of the vertex based on maximizing local mutual information. A way to reduce computational requirements of designating graph embeddings was proposed by Zhang et al. (2019). Xu et al. (2020) introduced GraphHeat which computed embeddings of vertices based on recognition of the local structure of their neighborhood.

Li et al. (2020) proposed to embed pairs of vertices instead of single vertices. But the resulting embeddings are designed to be easily translatable to embeddings of vertices.

c) *Graph autoencoders.*: These architectures implement the general structure of Variational Autoencoder (VAE, Kingma & Welling, 2014) and establish transformations  $\text{input\_graph} \rightarrow \text{embedding}$  (encoder) and  $\text{embedding} \rightarrow \text{input\_graph}$  (decoder). The embedding is in  $\mathbb{R}^{n \times d}$ , as discussed above.

Kipf & Welling (2016) proposed Graph VAE with an encoder in the form of graph convolutional NN and the decoder in the form of a simple inner product. Ma et al. (2018) introduced regularization to graph VAE that made the resulting graph satisfy constraints (e.g., imposed on molecules). Pan et al. (2018) proposed adversarial regularization for the embeddings to preserve the topological structure of the graph.

Accuracy of graph reproduction is the main problem in graph VAEs. Grover et al. (2019) introduced *Graphite*, a graph VAE with a strategy, inspired by low-rank approximations, of graph refinement in its decoder. Park et al. (2019) proposed an autoencoder with graph convolution, Laplacian smoothing of the encoder, and Laplacian sharpening-based decoder.

Attention (Vaswani et al., 2017) was introduced to graph VAE by Salehi & Davulcu (2019) as a crucial component of both the encoder and the decoder. Khan & Kleinstueber (2021) proposed a graph VAE aiming to maximize the similarity between the embeddings of neighboring and more distant vertices while minimizing the redundancy between the components of these embeddings.

d) *Graph generation.*: There are two generic approaches to graph generation, one based on Generative Adversarial Networks (GAN, Goodfellow et al., 2014) and one based on a sequential expansion of the graph.

In NetGAN (Bojchevski et al., 2018), the adjacency matrix is generated by a biased random walk among the vertices of the graph; the discriminator is an LSTM network that verifies if a walk through the graph is realistic. In MolGAN (De Cao & Kipf, 2018), the generator is a fixed size feedforward NN which produces an adjacency tensor; the discriminator is a graph convolutional NN.

---

#### Algorithm 1 Encoder

---

```

1: Input: Adjacency matrix  $A$ 
2: for  $k = 1, \dots, n - 1$  do
3:    $x_{k+1,k} \leftarrow e(\text{null}, \text{null}, A_{k+1,k})$ 
4: end for
5: for  $i = 1, \dots, n - 1$  do
6:   for  $k = 1, \dots, n - i$  do
7:      $x_{i+k+1,k} \leftarrow e(x_{i+k,k}, x_{i+k+1,k+1}, A_{i+k+1,k})$ 
8:   end for
9: end for
10: return  $x_{n,1}$  // embedding of input graph

```

---

Li et al. (2018) proposed a model that learns to assign probabilities to different actions within sequential graph generation, such as adding a vertex, adding an edge, etc. In GraphRNN (You et al., 2018), a recurrent NN is employed to fill in the adjacency matrix. Liao et al. (2019) used a recurrent NN with attention for the same purpose.

e) *Graph classification.*: A model able to learn to assign labels to graphs is a convolutional neural network (Defferrard et al., 2016). Zhang et al. (2018) added the SortPooling layer to such a network which enables its connection to a traditional NN as an output module. Errica et al. (2020) evaluated the above classifiers and others in an extensive study.

### III. METHOD

We define a **graph embedding** as a vector in  $\mathbb{R}^m$ , where  $m \in \mathbb{N}$  is an even constant. In the course of encoding a graph, embeddings,  $x_{i,j} \in \mathbb{R}^m$ ,  $i > j$ , are produced that approximately represent subgraphs of the given graph limited to its nodes  $j, \dots, i$ . These embeddings are produced recursively and finally,  $x_{n,1}$  represents the whole input graph. The structure of this recursion is presented in the left-hand part of Fig. 1 in reference to the adjacency matrix.

In the process of decoding the graph, embeddings,  $y_{i,j} \in \mathbb{R}^m$ , are recursively produced that represent subgraphs. Eventually, they represent single nodes.  $y$  has its specific indexing because decoding starts from the left lower corner of the adjacency matrix, whose size is initially unknown. In this specific indexing the coordinates of this left lower corner are  $\langle 0, 0 \rangle$ . The decoder recursion is presented in the middle part of Fig. 1. The right-hand part of this figure presents retrieving the original indexing of the adjacency matrix.

The **encoder** is a transformation

$$e : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \mapsto \mathbb{R}^m. \quad (1)$$

To produce  $x_{i,j}$ , the encoder is fed with (i) the embedding  $x_{i,j+1}$ , (ii) the embedding  $x_{i-1,j}$ , (iii) the entry  $A_{i,j}$ . Based on embeddings of two subgraphs, it produces an embedding that represents the union of these subgraphs. Applied recursively according to Algorithm 1, it finally produces  $x_{n,1}$  which represents the entire graph.

The **decoder** is a transformation

$$d : \mathbb{R}^m \mapsto \mathbb{R}^{m/2} \times \mathbb{R}^{m/2} \times \mathbb{R} \times \mathbb{R}. \quad (2)$$

Applied recursively according to Algorithm 2, it reconstructs the adjacency matrix. The input of  $d$  is an embedding,  $y_{i,j}$ , of the

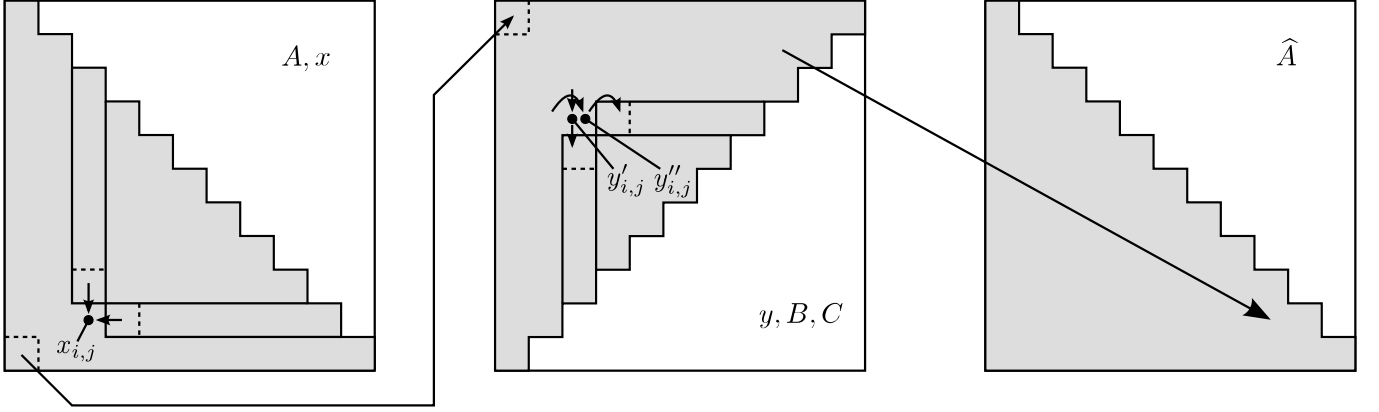


Fig. 1. Recursive graph autoencoder. *Left:* The encoder combines embeddings of smaller subgraphs into the embedding of their union. *Middle:* The decoder; for a given embedding of a subgraph, it produces half-embeddings for two smaller subgraphs, also an entry in a reindexed adjacency matrix,  $B$ , and an entry in a matrix,  $C$ , which indicates if the main diagonal of  $B$  has been reached. *Right:* By changing indexing in  $y$  we obtain  $\hat{A}$  which is a reconstruction of  $A$ .

---

**Algorithm 2** Decoder
 

---

```

1: Input: Graph embedding  $x$ 
2:  $y_{0,0} \leftarrow x$ 
3: for  $i = 0, \dots$  do
4:   for  $k = 0, \dots, i$  do
5:      $\langle y'_{i+1-k,k}, y''_{i-k,k+1}, B_{i-k,k}, C_{i-k,k} \rangle$ 
        $\leftarrow d(y_{i-k,k})$ 
6:   end for
7:    $y'_{0,i+1} \leftarrow d'(y_{0,i})$ 
8:    $y''_{i+1,0} \leftarrow d''(y_{i,0})$ 
9:   for  $k = 0, \dots, i+1$  do
10:     $y_{i+1-k,k} \leftarrow \text{concatenate}(y'_{i+1-k,k}, y''_{i+1-k,k})$ 
11:   end for
12:   if the average  $C_{i+1-k,k}$  for  $k \in \{0, \dots, i+1\}$  is below 0.5 then
13:     set  $n \leftarrow i+2$  and exit the loop.
14:   end if
15: end for
16: for  $i = 0, \dots, n-1; j = 0, \dots, n-1-i$  do
17:    $\hat{A}_{n-i,j+1} \leftarrow B_{i,j}$ 
18: end for
19: return  $\hat{A}$  // adjacency matrix estimate
  
```

---

subgraph of the target graph limited to its nodes  $j+1, \dots, n-i$ . When fed with  $y_{i,j}$ ,  $d$  produces (i) the left-hand half of the embedding  $y_{i+1,j}$ , (ii) the right-hand half of the embedding  $y_{i,j+1}$ , (iii) the value  $B_{i,j}$  that later on becomes the entry  $\hat{A}_{n-i,j+1}$  of the resulting adjacency matrix estimate, (iv) the value  $C_{i,j}$  that indicates if  $i+j+1 = n$ , i.e., if  $(i,j)$  has reached the antidiagonal of the  $B$  matrix (see the middle part of Figure 1).

*a) Training.* When training the encoder and decoder we require reconstruction of the input graph and a matrix,  $C$ , with ones at appropriate entries, thereby indicating the size of the output graph. That is, we require that

$$B_{i,j} = A_{n-i,j+1}, \quad C_{i,j} = 1 \quad (3)$$

for  $i+j = 0, \dots, n-2$ , and

$$B_{i,n-1-i} = 0, \quad C_{i,n-1-i} = 0 \quad (4)$$

for  $i = 0, \dots, n-1$ .

To facilitate the training on large graphs (with thousands of vertices), we utilize the fact that the encoder transforms subgraphs into embeddings, and the decoder transforms embeddings back into the same subgraphs. Therefore, we use subgraphs as training samples for the autoencoder. In subsequent epochs of the training the sizes of these subgraphs grow. We start with short recursion paths and gradually increase their lengths. Details of this procedure are presented in the appendix.

*b) Loss function.* A single graph contributes to the training loss with (i) a sum of cross-entropies for elements of  $A$  equal to 1, (ii) a sum of cross-entropies for elements of  $A$  equal to 0 and (iii) a square of the embedding norm. These components have their constant weights that assure that their contribution is comparable.

*c) Structure of encoder and decoder.* Our encoder and decoder are inspired by the GRU network (Cho et al., 2014). The encoder is designed to combine input embeddings of smaller subgraphs to the output embedding of their union. This combination is based on weighted averaging and additive adjusting. The encoder is based on a feedforward neural network  $f^e$  with a linear output layer, which produces three vectors of size  $m$ :

$$\langle z^0, z^1, \hat{x} \rangle = f^e(x^0, x^1, a). \quad (5)$$

The vectors  $z^0$  and  $z^1$  are applied for weighting input embeddings and  $\hat{x}$  is applied for additive adjusting. The output of the encoder is defined as

$$e(x^0, x^1, a) = (x^0 \circ \sigma(z^0) + x^1 \circ (\mathbf{1} - \sigma(z^0))) \circ \sigma(z^1) + \psi(\hat{x}) \circ (\mathbf{1} - \sigma(z^2)), \quad (6)$$

where  $\sigma$  is the logistic sigmoid,  $\psi$  is an activation function, and “ $\circ$ ” denotes the elementwise product.

The decoder is slightly more complex than the encoder, as it needs to combine inputs from two sides and produce outputs

in two directions (see the middle part of Fig. 1). The decoder processes an embedding, composed of two half-embeddings, of a subgraph and produces two half-embeddings of two smaller subgraphs along with an entry to the adjacency matrix and a marker of reaching the diagonal of this matrix. The output half-embeddings are produced with weighted averaging and additively adjusting the input half-embeddings. The decoder is based on a feedforward neural network,  $f^d$ , with a linear output layer which produces four vectors of size  $m/2$  and two scalars:

$$\langle z', z'', \hat{y}', \hat{y}'', b, c \rangle = f^d(\langle y', y'' \rangle). \quad (7)$$

They are applied to produce the output of the decoder as follows:

$$\begin{aligned} d(\langle y', y'' \rangle) \\ = \langle y' \circ \sigma(z') + \psi(\hat{y}') \circ (\mathbf{1} - \sigma(z')), \\ y'' \circ \sigma(z'') + \psi(\hat{y}'') \circ (\mathbf{1} - \sigma(z'')), b, c \rangle. \end{aligned} \quad (8)$$

For the upper row and the leftmost column of the  $y$  matrix, there are no half-embeddings that come from above and from the left, respectively. Therefore, we use two additional networks,  $f^{d1}$  and  $f^{d2}$  to produce these embeddings. Those networks proceed as follows:

$$\langle z', \hat{y}' \rangle = f^{d1}(\langle y', y'' \rangle) \quad (9)$$

$$\langle z'', \hat{y}'' \rangle = f^{d2}(\langle y', y'' \rangle) \quad (10)$$

The first half-embeddings for the upper row of  $y$  and the second half-embeddings for the leftmost columns are produced, respectively, as follows

$$d'(\langle y', y'' \rangle) = y' \circ \sigma(z') + \psi(\hat{y}') \circ (\mathbf{1} - \sigma(z')), \quad (11)$$

$$d''(\langle y', y'' \rangle) = y'' \circ \sigma(z'') + \psi(\hat{y}'') \circ (\mathbf{1} - \sigma(z'')). \quad (12)$$

#### A. Order of vertices and adjacency matrix patches

We apply two known techniques to boost the autoencoder efficiency (Salehi & Davulcu, 2019). Since indexing of vertices is arbitrary, we proceed as follows: The vertices are sorted by their degree in decreasing order. Then, breadth-first search (BFS) runs in the graph starting from the first vertex. Order of occurring of vertices in BFS defines their indexing for the autoencoder.

Our basic model operates on single entries in the adjacency matrix. However, it may also operate on  $l \times l$  patches of this matrix for  $l \in \mathbb{N}$ . This way, recursion length is reduced  $l$  times, and inputs/outputs of the encoder/decoder become  $l \times l$  matrices rather than scalars. The entries in the patches of  $A$  that reach their diagonal and further are assumed equal to -1. The corresponding entries in the patches of  $C$  are required to be 0. We apply this extension in the experimental study below.

#### B. Variational Graph Autoencoder

In order to extend our proposed graph autoencoder to the graph VAE, we add two elements:

- A feedforward neural network that transforms graph embeddings,  $x$ , into vectors,  $\rho \in \mathbb{R}^m$ , of logarithms of standard deviations.

- KL-divergence between  $\mathcal{N}(x, \exp(\rho))$  and  $\mathcal{N}(0, I)$  as a part of the training loss.

As in the original VAE (Cho et al., 2014), when training the graph VAE, the decoder is fed with  $x + \xi \circ \exp(\rho)$ ,  $\xi \sim \mathcal{N}(0, I)$ .

#### C. Further extensions

Graphs representing complex systems often have numerical information assigned to their vertices and edges. Our architecture can readily be extended to model and process this information:

- **Weighted edges:** The weights are inputs to the encoder and outputs of the decoder combined with the adjacency matrix entries.
- **Labeled edges:** Related to the previous point. The input/output to the encoder/decoder may contain other information than just the graph structure. By treating the entries as vectors instead of scalars, our method could process graphs with labeled edges.
- **Labeled vertices:** Auxiliary feedforward networks transform the labels into embeddings of single-vertex graphs and back. Together with the previous point, it would enable the method to work with graphs describing objects and their relationships.
- **Directed graphs:** The encoder is fed with, and the decoder produces, the pair  $\langle A_{i,j}, A_{j,i} \rangle$  instead of just  $A_{i,j}$ .

### IV. EXPERIMENTAL STUDY

There are two goals of this study. The first one is to evaluate our proposed autoencoder in reaching its primary purpose, i.e., is encoding and decoding graphs. The second one is to evaluate the graph embeddings it produces as input to conventional classifiers.

#### A. Datasets

We use six sets of social graphs from (Yanardag & Vishwanathan, 2015b): IMDB-BINARY, IMDB-MULTI, COLLAB, REDDIT-BINARY, REDDIT-MULTI-5K, REDDIT-MULTI-12K and GRID-MEDIUM – a set of synthetic grid-like graphs. Their basic statistics are contained in Table I.

The datasets are divided 70/15/15 into training, validation, and test subsets. These subsets are augmented by  $N$  random permutations of each graph.  $N = 99$  for GRID-MEDIUM;  $N = 9$  for IMDB-BINARY, IMDB-MULTI and COLLAB;  $N = 0$  for all REDDITs.

#### B. Graph autoencoding

We have found two difficulties in comparing our proposed autoencoder with the state-of-the-art. Firstly, existing autoencoders are based on embeddings whose size depends on the graph size. Secondly, they usually learn on a single graph, rather than on a set of graphs. A set of graphs can be understood as one graph with a number of separate components. However, an autoencoder trained on such a supergraph is not necessarily applicable to its hypothetical test components. Autoencoders that we have found technically capable of being trained on a set of graphs and tested on separate graphs are GAE and

Dataset	Size	Avg. Num Nodes	Max. Num Nodes	Avg. Num Edges	Fill	Num Classes
GRID-MEDIUM	49	25.0	64	40.0	0.18	-
IMDB-BINARY	1000	19.8	136	96.5	0.52	92
IMDB-MULTI	1500	13.0	89	65.9	0.77	3
COLLAB	5000	74.5	492	2457.5	0.51	3
REDDIT-BINARY	2000	429.6	3782	497.8	0.02	2
REDDIT-MULTI-5K	4999	508.5	3648	594.9	0.01	5
REDDIT-MULTI-12K	11929	391.4	3782	456.9	0.02	11

TABLE I

PROPERTIES OF THE DATASETS USED IN OUR EXPERIMENTS. *Size* IS THE NUMBER OF GRAPHS IN THE DATASET, *Fill* IS THE AVERAGE PROPORTION OF THE NUMBER OF EDGES TO THE NUMBER OF ENTRIES IN THE ADJACENCY MATRICES.

VGAE (Kipf & Welling, 2016). The goal of experiments is to learn to encode and decode graphs on the training part of the dataset. The performance of these operations is verified on the test part of the dataset.

When applying GAE and VGAE, we used a vertex embedding size,  $d$ , equal to 4 for REDDIT datasets, and equal to 8 for other datasets. The graph embedding size in our proposed autoencoder was equal to  $d$  times the average number of vertices for graphs in the dataset. Consequently, the graph embedding sizes were on average equal for different methods. These settings gave GAE and VGAE an advantage as these algorithms could embed relatively large graphs in accordingly large vectors.

All architectures were trained with early stopping and a patience of 20 epochs. We have designated hyperparameters of all methods and datasets with random search and on a number of evaluations specific to the dataset. Values of those hyperparameters are presented in the appendix.

All our experiments are repeated 5 times with different random seeds and on different pre-prepared data splits. We report averaged results.

Tab. III presents precision and recall, and Tab. II presents the F1 score for the reconstruction of ones in the adjacency matrix. These statistics were calculated as follows: For each graph, they were averaged over all entries of the adjacency matrix, and then over the test graphs with a weight equal to the number of vertices within a graph. Compared to GAE and VGAE, our method yields comparable or better results. On the bigger, more difficult datasets, especially the REDDIT graphs, our method is incomparably more accurate. Neither GAE nor VGAE could be trained to provide meaningful output for the harder datasets, often generating only 0s or 1s. Our method is more capable of recognizing the graph structure, as evidenced by the balanced recall and precision scores on every dataset. Tab. III details the related recall and precision scores.

Fig. 2 presents examples of reconstructed graphs.

Contrary to GAE and VGAE, our autoencoder also learns to properly encode the sizes of graphs, which implies it can make an error related to graph size. As evident in Tab. IV, such errors are generally small in magnitude. For the calculation of edge recall and precision scores, the graphs of incorrect sizes (either predicted or target) were padded with 0s along the adjacency matrix diagonal, which serves as an added penalty.

### C. Graph classification

To reach the second goal of this study, we use embeddings of graphs analyzed in the previous subsection and feed them

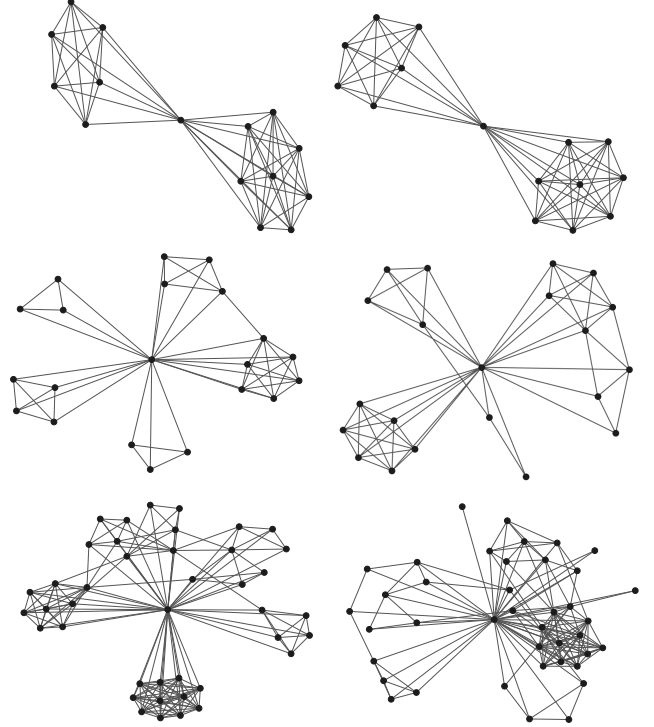


Fig. 2. Comparison of selected accurate or slightly flawed graph reconstruction results on the IMDB datasets. Original graphs are on the left and the reconstructed ones are on the right.

to popular classifiers such as SVM, random forests, and others. For an ablation, we train the same classifiers with vectors of basic statistics describing the graphs. These statistics are: (i) the number of nodes, (ii) the average vertex degree, (iii) the root of the mean square of vertex degree, and (iv) the square of the average root of vertex degree.

Tab. VI presents the comparison of the accuracy of SVM trained on the embeddings and SVM trained on the basic statistics. Results of classifiers other than SVM are presented in the appendix; they are similar. It is seen that shallow processing of the embeddings does not lead to accurate classification. We consider this result negative. However, according to Errica et al. (2020), only a few of the well established methods outperform our naive classifier.

## V. CONCLUSIONS

In this paper, we have introduced a recursive neural architecture capable of representing graphs of any size in vectors of fixed dimension (embeddings), and capable of reconstructing

Dataset\method	GAE	VGAE	Ours
GRID-MEDIUM	$0.46 \pm 0.10$	$0.43 \pm 0.09$	$0.70 \pm 0.22$
IMDB-BINARY	$0.91 \pm 0.02$	$0.86 \pm 0.02$	$0.90 \pm 0.02$
IMDB-MULTI	$0.93 \pm 0.01$	$0.89 \pm 0.01$	$0.89 \pm 0.01$
COLLAB	$0.78 \pm 0.01$	$0.75 \pm 0.02$	$0.86 \pm 0.01$
REDDIT-BINARY	$0.02 \pm 0.00$	$0.01 \pm 0.00$	$0.53 \pm 0.02$
REDDIT-MULTI-5K	$0.01 \pm 0.00$	$0.01 \pm 0.00$	$0.48 \pm 0.01$
REDDIT-MULTI-12K	$0.01 \pm 0.00$	$0.01 \pm 0.00$	$0.49 \pm 0.01$

TABLE II

AUTOENCODERS: F1 SCORE ON THE TEST SET.

Dataset\method	GAE		VGAE		Ours	
	Precision	Recall	Precision	Recall	Precision	Recall
GRID-MEDIUM	$0.31 \pm 0.08$	$0.98 \pm 0.00$	$0.28 \pm 0.07$	$0.97 \pm 0.01$	$0.69 \pm 0.24$	$0.71 \pm 0.20$
IMDB-BINARY	$0.88 \pm 0.03$	$0.95 \pm 0.01$	$0.82 \pm 0.03$	$0.91 \pm 0.00$	$0.87 \pm 0.02$	$0.93 \pm 0.02$
IMDB-MULTI	$0.90 \pm 0.02$	$0.95 \pm 0.01$	$0.87 \pm 0.02$	$0.91 \pm 0.01$	$0.86 \pm 0.01$	$0.92 \pm 0.01$
COLLAB	$0.65 \pm 0.01$	$0.96 \pm 0.00$	$0.62 \pm 0.02$	$0.95 \pm 0.00$	$0.89 \pm 0.01$	$0.83 \pm 0.02$
REDDIT-BINARY	$0.01 \pm 0.00$	$0.65 \pm 0.09$	$0.01 \pm 0.00$	$0.97 \pm 0.05$	$0.59 \pm 0.02$	$0.48 \pm 0.02$
REDDIT-MULTI-5K	$0.01 \pm 0.00$	$0.75 \pm 0.05$	$0.00 \pm 0.00$	$0.98 \pm 0.03$	$0.54 \pm 0.03$	$0.43 \pm 0.01$
REDDIT-MULTI-12K	$0.01 \pm 0.00$	$0.76 \pm 0.07$	$0.01 \pm 0.00$	$0.97 \pm 0.02$	$0.53 \pm 0.05$	$0.45 \pm 0.03$

TABLE III

AUTOENCODERS: PRECISION AND RECALL ON THE TEST SET.

Dataset	Size accuracy	Mean size error
GRID-MEDIUM	$0.34 \pm 0.11$	$0.100 \pm 0.039$
IMDB-BINARY	$0.97 \pm 0.03$	$0.003 \pm 0.002$
IMDB-MULTI	$0.97 \pm 0.01$	$0.003 \pm 0.002$
COLLAB	$0.97 \pm 0.01$	$0.001 \pm 0.001$
REDDIT-BINARY	$0.23 \pm 0.04$	$0.029 \pm 0.010$
REDDIT-MULTI-5K	$0.38 \pm 0.10$	$0.006 \pm 0.002$
REDDIT-MULTI-12K	$0.54 \pm 0.15$	$0.021 \pm 0.018$

TABLE IV

SIZE ERRORS: SIZE ACCURACY — THE NUMBER OF GRAPHS WITH CORRECTLY DESIGNATED VERTEX COUNT OVER THE NUMBER OF ALL GRAPHS, MEAN SIZE ERROR — THE AVERAGE ABSOLUTE SIZE DIFFERENCE BETWEEN THE TARGET AND PREDICTED GRAPHS OVER THE AVERAGE TARGET GRAPH SIZE.

Dataset	basic stats.	embeddings
IMDB-BINARY	$0.72 \pm 0.01$	$0.84 \pm 0.01$
IMDB-MULTI	$0.51 \pm 0.01$	$0.71 \pm 0.01$
COLLAB	$0.70 \pm 0.00$	$0.86 \pm 0.02$
REDDIT-BINARY	$0.81 \pm 0.01$	$0.78 \pm 0.01$
REDDIT-MULTI-5K	$0.54 \pm 0.01$	$0.49 \pm 0.01$
REDDIT-MULTI-12K	$0.43 \pm 0.00$	$0.40 \pm 0.02$

TABLE V

CLASSIFIERS: ACCURACY ON THE TRAIN SET WITH SVM CLASSIFIER.

these graphs from the vectors. The architecture does not impose any structural upper bound on the size of the graph or lower bound on the dimension of the embeddings. In our experiments, we verified the proposed method on 7 datasets with graphs with up to 3782 vertices. Given these experiments, our architecture is effective and offers a reasonable level of accuracy.

Fixed-size embeddings of graphs enable a wide range of further developments such as graph generation, completion,

Dataset	basic stats.	embeddings
IMDB-BINARY	$0.70 \pm 0.02$	$0.64 \pm 0.01$
IMDB-MULTI	$0.47 \pm 0.03$	$0.47 \pm 0.02$
COLLAB	$0.71 \pm 0.01$	$0.75 \pm 0.01$
REDDIT-BINARY	$0.81 \pm 0.03$	$0.80 \pm 0.04$
REDDIT-MULTI-5K	$0.54 \pm 0.02$	$0.46 \pm 0.01$
REDDIT-MULTI-12K	$0.43 \pm 0.00$	$0.39 \pm 0.01$

TABLE VI

CLASSIFIERS: ACCURACY ON THE TEST SET WITH SVM CLASSIFIER.

or transformation. These may lead to applications that are impossible at the moment.

## REFERENCES

- Bai, Y., Ding, H., Qiao, Y., Marinovic, A., Gu, K., Chen, T., Sun, Y., and Wang, W. Unsupervised inductive graph-level representation learning via graph-graph proximity. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1988–1994, 2019.
- Bojchevski, A., Shchur, O., Zügner, D., and Günnemann, S. Netgan: Generating graphs via random walks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 609–618, 2018.
- Cao, S., Lu, W., and Xu, Q. Deep neural networks for learning graph representations. In *Proceedings of the 30-th AAAI Conference on Artificial Intelligence*, pp. 1145–1152, 2016.
- Chami, I., Ying, Z., Ré, C., and Leskovec, J. Hyperbolic graph convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)*, volume 32, pp. 4868–4879, 2019.
- Cho, K., Merriënboer, B. V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP2014)*, 2014.
- Clevert, D., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1511.07289>.
- Dai, H., Kozareva, Z., Dai, B., Smola, A., and Song, L. Learning steady-states of iterative algorithms over graphs. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1106–1114, 2018.
- De Cao, N. and Kipf, T. Molgan: An implicit generative model for small molecular graphs, 2018. arXiv:1805.11973.

- Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- Errica, F., Podda, M., Bacciu, D., and Micheli, A. A fair comparison of graph neural networks for graph classification. In *International Conference on Learning Representations (ICLR)*, 2020.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Networks. In *Neural Information Processing Systems (NIPS)*, 2014.
- Grover, A., Zweig, A., and Ermon, S. Graphite: Iterative generative modeling of graphs. In *International conference on machine learning (ICML)*, pp. 2434–2444, 2019.
- Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in neural information processing systems (NIPS)*, pp. 1024–1034, 2017.
- Khan, R. A. and Kleinstuber, M. Barlow graph auto-encoder for unsupervised network embedding, 2021. arXiv:2110.15742.
- Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.
- Kipf, T. and Welling, M. Variational graph auto-encoders, 2016. arXiv:1611.07308.
- Klicpera, J., Weißenberger, S., and Günnemann, S. Diffusion improves graph learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13354–13366, 2019.
- Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. W. Learning deep generative models of graphs, 2018. arXiv:1803.03324.
- Li, Y., Luo, B., and Gui, N. Pair-view unsupervised graph representation learning, 2020. arXiv:2012.06113.
- Liao, R., Li, Y., Song, Y., Wang, S., Hamilton, W., Duvenaud, D. K., Urtasun, R., and Zemel, R. Efficient graph generation with graph recurrent attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.
- Ma, T., Chen, J., and Xiao, C. Constrained generation of semantically valid graphs via regularizing variational autoencoders. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7113–7124, 2018.
- Mehta, N., Duke, L. C., and Rai, P. Stochastic blockmodels meet graph neural networks. In *International Conference on Machine Learning (ICML)*, pp. 4466–4474, 2019.
- Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., and Jaiswal, S. graph2vec: Learning distributed representations of graphs. In *International Workshop on Mining and Learning with Graphs (KDD MLG)*, 2020.
- Nowicki, K. and Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.
- Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., and Zhang, C. Adversarially regularized graph autoencoder for graph embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2609–2615, 2018.
- Park, J., Lee, M., Chang, H. J., Lee, K., and Choi, J. Y. Symmetric graph convolutional autoencoder for unsupervised graph representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6519–6528, 2019.
- Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond, 2019.
- Salehi, A. and Davulcu, H. Graph attention auto-encoders, 2019. arXiv:1905.10715.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems (NIPS)*, 2017.
- Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. Deep graph infomax. In *International Conference on Learning Representations (ICLR)*, 2019.
- Wang, D., Cui, P., and Zhu, W. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1225–1234, 2016.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32: 4–24, 2019.
- Xu, B., Shen, H., Cao, Q., Cen, K., and Cheng, X. Graph convolutional networks using heat kernel for semi-supervised learning, 2020. arXiv:2007.16002.
- Yanardag, P. and Vishwanathan, S. Deep graph kernels. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1365–1374, 2015a.
- Yanardag, P. and Vishwanathan, S. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1365–1374, 2015b.
- Yang, Z., Cohen, W. W., and Salakhutdinov, R. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, pp. 40–48, 2016.
- Ying, R., You, J., Morris, C., Ren, X., Hamilton, W. L., and Leskovec, J. Hierarchical graph representation learning with differentiable pooling. In *Neural Information Processing Systems (NeurIPS)*, 2018.
- You, J., Ying, R., Ren, X., Hamilton, W. L., and Leskovec, J. Graphrnn: Generating realistic graphs with deep autoregressive models. In *Proceedings of the 35th International Conference on International Conference on Machine Learning (ICML)*, pp. 5694–5703, 2018.
- Zhang, J., Dong, Y., Wang, Y., Tang, J., and Ding, M. ProNE: fast and scalable network representation learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4278–4284, 2019.
- Zhang, M., Cui, Z., Neumann, M., and Chen, Y. An end-to-end deep learning architecture for graph classification. In *Proceedings of the 32-nd AAAI Conference on Artificial Intelligence*, 2018.

Zhang, X. and Chen, L. Capsule graph neural network. In *International Conference on Learning Representations (ICLR)*, 2019.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. Graph neural networks: A review of methods and applications, 2019. arxiv:1812.08434v4.

#### APPENDIX

For each algorithm-dataset pair, we have determined approximately optimal hyperparameters in the same procedure based on a random search and a limited grid search.

For each dataset-algorithm pair, we have defined the space for the random search with the upper bound three orders of magnitude above the lower bound. These bounds were determined based on the hyperparameters’ values that had successfully worked in analogical dataset-algorithm pairs. We launched 50 runs of random search to designate good values for further evaluation. If the search ranges turned out to be misaligned, the second set of 50 runs with narrowed-down ranges produced competitive results. Then, the best hyperparameters in the grid were identified. For each hyperparameter, there were only two values in the grid, i.e., the smaller and the greater than the value resulting from the random search.

The networks generally achieve slightly better performance with the hyperparameters designated with the random search than with the final tuning, but the difference is usually negligible.

Hyperparameters of GAE/VGAE are presented in Tab. VII. Hyperparameters of our proposed autoencoder are presented in Tab. VIII.

We use the standard algorithms based on stochastic gradient descent to optimize the weights of the neural networks. In particular, we have found that the AdamW Loshchilov & Hutter (2019) optimizer in the AMSGrad variant Reddi et al. (2019) produces the best results on our model.

Training on graphs larger than about 1000 nodes proves difficult for the autoencoder at the beginning of the training. To alleviate this problem, we introduce a custom training procedure called “progressive subgraph training”, which involves training the network to reconstruct smaller subgraphs of the original graphs. The subgraphs are constructed by splitting the adjacency matrix into smaller triangular blocks ending on the main diagonal. Fig. 3 demonstrates the idea behind their shape. The subgraphs are temporarily treated as the training dataset for the autoencoder so that the model learns to reconstruct them. At first, the subgraphs are small and get larger as the procedure advances. We determine when to proceed to the larger subgraphs once the network’s recall and precision performance exceed a threshold of 0.1 each. Once the subgraphs reach the size of the original graphs, the model proceeds learning on the original dataset as normal.

The network’s behavior during the training has also shown that training on larger graphs is smoother when used with lower learning rates. Therefore, it proved advantageous to gradually reduce the learning rate by 10% of the current value at each increase in the size of subgraphs.

We utilize the progressive subgraph training procedure to train the network on the REDDIT datasets. There, we set the

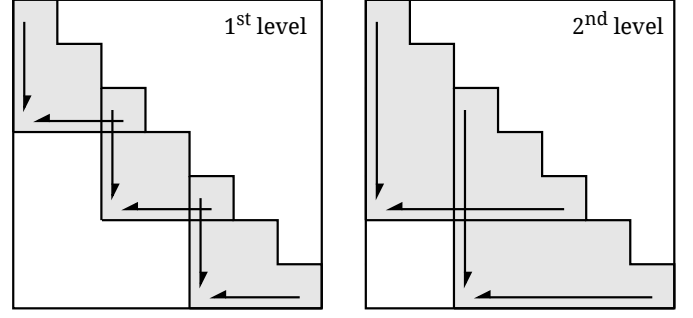


Fig. 3. Subgraphs as they are constructed during subsequent levels of progressive subgraph training. The levels and size proportions do not align with the actual used values, as they are depicted here for illustrative purposes.

initial size of the subgraphs at 5% of the original graph size and increase the size by 10% points of the maximum size at each level.

Results of classification of graphs based on basic statistics and embeddings is presented in Tab. IX and Tab. X, with performance on the training- and test sets, respectively. We used scikit learn implementation of models with default parameters. There was one exception only – in random forest we used `min_samples_leaf=20`, as it gave better results than the default value.

All our experiments were performed on two PCs, each equipped with Ryzen™ Threadripper™ 1920X and 4x GeForce™ RTX 2070 Super.

Our experimental software has been written in python 3.9 with pytorch-lightning 1.5.6. Upon acceptance of this paper we will publish our software.



Dataset	hidden	latent	lr	dropout	recall prec bias	batch
GRID-MEDIUM	512	8	0.005	0.1	0.5	32
IMDB-BINARY	512	8	0.005	0.1	0.5	32
IMDB-MULTI	512	8	0.005	0.1	0.5	32
COLLAB	4096	8	0.005	0.1	0.5	32
REDDIT-BINARY	4096	4	0.005	0.1	0.003	32
REDDIT-MULTI-5K	4096	4	0.005	0.1	0.003	32
REDDIT-MULTI-12K	4096	4	0.005	0.1	0.003	32

TABLE VII

HYPERPARAMETERS OF GAE/VGAE KIPF & WELLING (2016), THE SAME FOR BOTH: HIDDEN – HIDDEN LAYER SIZE, LATENT – LATENT SIZE, LR – LEARNING RATE, RECALL PREC BIAS – PROPORTION OF WEIGHTS OF TARGET 0s AND 1s IN LOSS FUNCTION, HIGHER VALUES FAVOR RECALL PERFORMANCE BY INCREASING THE WEIGHTS OF 1s, BATCH – BATCH SIZE.

Dataset	embed.	encoder	decoder	block	lr	grad. clip	recall prec bias	batch
GRID-MEDIUM	200	2048	2048	4	0.0003	1.0	0.3	32
IMDB-BINARY	160	1024:768	2048	4	0.0005	1.0	0.5	64
IMDB-MULTI	104	1024	2048	8	0.0005	1.0	0.5	64
COLLAB	604	2048:1536	4096	16	0.0003	0.5	0.3	32
REDDIT-BINARY	1720	4096	6144	64	0.0003	1.0	0.03	32
REDDIT-MULTI-5K	2036	4096	6144	64	0.0003	0.5	0.03	32
REDDIT-MULTI-12K	1564	4096	6144	64	0.0003	0.5	0.03	32

TABLE VIII

HYPERPARAMETERS OF OUR AUTOENCODER FOR THE RECONSTRUCTION TASK: EMBED. – EMBEDDING SIZE, ENCODER – SIZES OF ENCODER HIDDEN LAYERS, DECODER – SIZES OF DECODER HIDDEN LAYERS, BLOCK – BLOCK SIZE, LR – LEARNING RATE, GRAD. CLIP – GRADIENT CLIPPING VALUE, RECALL PREC BIAS – PROPORTION OF WEIGHTS OF TARGET 0s AND 1s IN LOSS FUNCTION, HIGHER VALUES FAVOR RECALL PERFORMANCE BY INCREASING THE WEIGHTS OF 1s, BATCH – BATCH SIZE. FOR ALL EXPERIMENTS WE SET 0.5 AS THE MASK WEIGHT AND 0.2 AS THE EMBEDDING NORM WEIGHT FOR THE LOSS CALCULATION. IN BOTH THE ENCODER AND THE DECODER WE USE ELU CLEVERT ET AL. (2016) AS THE ACTIVATION FUNCTION.

Dataset→ Method↓	IMDB- BINARY	IMDB- MULTI	COLLAB	REDDIT- BINARY	REDDIT- MULTI-5k	REDDIT- MULTI-12k
basic stats. + Naive Bayes	0.66 ± 0.01	0.40 ± 0.01	0.63 ± 0.01	0.68 ± 0.01	0.50 ± 0.01	0.32 ± 0.00
basic stats. + logistic regression	0.65 ± 0.01	0.39 ± 0.03	0.69 ± 0.00	0.71 ± 0.01	0.54 ± 0.00	0.39 ± 0.00
basic stats. + random forest	0.87 ± 0.01	0.72 ± 0.01	0.95 ± 0.00	0.90 ± 0.01	0.61 ± 0.01	0.52 ± 0.00
basic stats. + SVM	0.72 ± 0.01	0.51 ± 0.01	0.70 ± 0.00	0.81 ± 0.01	0.54 ± 0.01	0.43 ± 0.00
embeddings + Naive Bayes	0.65 ± 0.01	0.49 ± 0.03	0.61 ± 0.01	0.71 ± 0.03	0.42 ± 0.01	0.19 ± 0.01
embeddings + logistic regression	0.72 ± 0.01	0.58 ± 0.01	0.79 ± 0.01	0.83 ± 0.00	0.51 ± 0.01	0.41 ± 0.01
embeddings + random forest	0.88 ± 0.00	0.74 ± 0.02	0.90 ± 0.00	0.86 ± 0.01	0.65 ± 0.00	0.54 ± 0.01
embeddings + SVM	0.84 ± 0.01	0.71 ± 0.01	0.86 ± 0.02	0.78 ± 0.01	0.49 ± 0.01	0.40 ± 0.02

TABLE IX

CLASSIFIERS: FINAL ACCURACY ON THE TRAINING SET.

Dataset→ Method↓	IMDB- BINARY	IMDB- MULTI	COLLAB	REDDIT- BINARY	REDDIT- MULTI-5Kk	REDDIT- MULTI-12K
basic stats. + Naive Bayes	0.64 ± 0.02	0.39 ± 0.04	0.65 ± 0.01	0.67 ± 0.02	0.52 ± 0.01	0.32 ± 0.01
basic stats. + logistic regression	0.63 ± 0.05	0.35 ± 0.03	0.70 ± 0.01	0.70 ± 0.03	0.53 ± 0.01	0.39 ± 0.01
basic stats. + random forest	0.65 ± 0.02	0.46 ± 0.03	0.71 ± 0.02	0.87 ± 0.01	0.54 ± 0.02	0.44 ± 0.01
basic stats. + SVM	0.70 ± 0.02	0.47 ± 0.03	0.71 ± 0.01	0.81 ± 0.03	0.54 ± 0.02	0.43 ± 0.00
embeddings + Naive Bayes	0.61 ± 0.02	0.42 ± 0.06	0.60 ± 0.01	0.72 ± 0.01	0.41 ± 0.02	0.18 ± 0.01
embeddings + logistic regression	0.65 ± 0.03	0.45 ± 0.04	0.71 ± 0.00	0.82 ± 0.02	0.49 ± 0.01	0.40 ± 0.01
embeddings + random forest	0.68 ± 0.02	0.47 ± 0.02	0.74 ± 0.01	0.83 ± 0.03	0.47 ± 0.01	0.39 ± 0.01
embeddings + SVM	0.64 ± 0.01	0.47 ± 0.02	0.75 ± 0.01	0.80 ± 0.04	0.46 ± 0.01	0.39 ± 0.01

TABLE X

CLASSIFIERS: FINAL ACCURACY ON THE TEST SET.