

# Message Passing Attention Networks for Document Understanding

Giannis Nikolentzos,<sup>1</sup> Antoine J.-P. Tixier,<sup>1</sup> Michalis Vazirgiannis<sup>1,2</sup>

<sup>1</sup>École Polytechnique

<sup>2</sup>Athens University of Economics and Business  
{nikolentzos,anti5662,mvazirg}@lix.polytechnique.fr

## Abstract

Graph neural networks have recently emerged as a very effective framework for processing graph-structured data. These models have achieved state-of-the-art performance in many tasks. Most graph neural networks can be described in terms of message passing, vertex update, and readout functions. In this paper, we represent documents as word co-occurrence networks and propose an application of the message passing framework to NLP, the Message Passing Attention network for Document understanding (MPAD). We also propose several hierarchical variants of MPAD. Experiments conducted on 10 standard text classification datasets show that our architectures are competitive with the state-of-the-art. Ablation studies reveal further insights about the impact of the different components on performance. Code is publicly available at: <https://github.com/giannisnik/mpad>.

## 1 Introduction

The concept of message passing over graphs has been around for many years (Weisfeiler and Lehman 1968; Murphy, Weiss, and Jordan 1999), as well as that of graph neural networks (GNNs) (Gori, Monfardini, and Scarselli 2005; Scarselli et al. 2008). However, GNNs have only recently started to be closely investigated, following the advent of deep learning. Some notable examples include (Duvenaud et al. 2015; Battaglia et al. 2016; Li et al. 2016; Defferrard, Bresson, and Vandergheynst 2016; Kearnes et al. 2016; Kipf and Welling 2016; Hamilton, Ying, and Leskovec 2017; Veličković et al. 2017; Xu et al. 2018b). These approaches are known as *spectral*. Their similarity with message passing (MP) was observed by (Kipf and Welling 2016) and formalized by (Gilmer et al. 2017) and (Xu et al. 2018a).

The MP framework is based on the core idea of *recursive neighborhood aggregation*. That is, at every iteration, the representation of each vertex is updated based on messages received from its neighbors. The majority of the spectral GNNs can be described in terms of the MP framework.

GNNs have been applied with great success to bioinformatics and social network data, for node classification, link prediction, and graph classification. However, a few studies only have focused on the application of the MP framework to representation learning on text. This paper proposes one such application. More precisely, we represent documents

as word co-occurrence networks, and develop an expressive MP GNN tailored to document understanding, the Message Passing Attention network for Document understanding (MPAD). We also propose several hierarchical variants of MPAD. Evaluation on 10 document classification datasets shows that our architectures learn representations that are competitive with the state-of-the-art. Furthermore, ablation experiments shed light on the impact of various architectural choices.

In what follows, we first provide some background about the MP framework (sec. 2), thoroughly describe and explain MPAD (sec. 3), present our experimental framework (sec. 4), report and interpret our results (sec. 5), and provide a review of the relevant literature (sec. 6).

## 2 Message Passing Neural Networks

(Gilmer et al. 2017) proposed a MP framework under which many of the recently introduced GNNs can be reformulated<sup>1</sup>. MP consists in an aggregation phase followed by a combination phase (Xu et al. 2018a). More precisely, let  $G = (V, E)$  be a graph, and let us consider  $v \in V$ . At time  $t + 1$ , a message vector  $\mathbf{m}_v^{t+1}$  is computed from the representations of the neighbors  $\mathcal{N}(v)$  of  $v$ :

$$\mathbf{m}_v^{t+1} = \text{AGGREGATE}^{t+1}(\{\mathbf{h}_w^t \mid w \in \mathcal{N}(v)\}) \quad (1)$$

The new representation  $\mathbf{h}_v^{t+1}$  of  $v$  is then computed by combining its current feature vector  $\mathbf{h}_v^t$  with the message vector  $\mathbf{m}_v^{t+1}$ :

$$\mathbf{h}_v^{t+1} = \text{COMBINE}^{t+1}(\mathbf{h}_v^t, \mathbf{m}_v^{t+1}) \quad (2)$$

Messages are passed for  $T$  time steps. Each step is implemented by a different layer of the MP network. Hence, iterations correspond to network depth. The final feature vector  $\mathbf{h}_v^T$  of  $v$  is based on messages propagated from all the nodes in the subtree of height  $T$  rooted at  $v$ . It captures both the topology of the neighborhood of  $v$  and the distribution of the vertex representations in it.

If a graph-level feature vector is needed, e.g., for classification or regression, a READOUT pooling function, that

<sup>1</sup>Note that some GNNs, known as *spatial*, are not based on MP (Niepert, Ahmed, and Kutzkov 2016; Nikolentzos et al. 2018; Tixier et al. 2019).

must be invariant to permutations, is applied:

$$\mathbf{h}_G = \text{READOUT}(\{\mathbf{h}_v^T \mid v \in V\}) \quad (3)$$

Next, we present the MP network we developed for document understanding.

### 3 Message Passing Attention network for Document understanding (MPAD)

#### 3.1 Word co-occurrence networks

We represent a document as a statistical word co-occurrence network (Mihalcea and Tarau 2004) with a sliding window of size 2 overspanning sentences. Let us denote that graph by  $G = (V, E)$ . Each unique word in the preprocessed document is represented by a node in  $G$ , and an edge is added between two nodes if they are found together in at least one instantiation of the window.  $G$  is directed and weighted: edge directions and weights respectively capture text flow and co-occurrence counts.

$G$  is a compact representation of its document. In  $G$ , immediate neighbors are consecutive words in the same sentence<sup>2</sup>. That is, paths of length 2 correspond to bi-grams. Paths of length more than 2 can correspond either to traditional  $n$ -grams or to *relaxed*  $n$ -grams, that is, words that never appear in the same sentence but co-occur with the same word(s). Such nodes are linked through common neighbors.

**Master node.** Inspired by (Scarselli et al. 2008), our graph  $G$  also includes a special document node, linked to all other nodes via unit weight bi-directional edges. In what follows, let us denote by  $n$  the number of nodes in  $G$ , including the master node.

#### 3.2 Message passing

We formulate our AGGREGATE function as:

$$\mathbf{M}^{t+1} = \text{MLP}^{t+1}(\mathbf{D}^{-1}\mathbf{A}\mathbf{H}^t) \quad (4)$$

where  $\mathbf{H}^t \in \mathbb{R}^{n \times d}$  contains node features ( $d$  is a hyperparameter<sup>3</sup>), and  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is the adjacency matrix of  $G$ . Since  $G$  is directed,  $\mathbf{A}$  is asymmetric. Also,  $\mathbf{A}$  has zero diagonal as we choose not to consider the feature of the node itself, only that of its incoming neighbors, when updating its representation<sup>4</sup>. Since  $G$  is weighted, the  $i^{th}$  row of  $\mathbf{A}$  contains the weights of the edges incoming on node  $v_i$ .  $\mathbf{D} \in \mathbb{R}^{n \times n}$  is the diagonal in-degree matrix of  $G$ . MLP denotes a multi-layer perceptron, and  $\mathbf{M}^{t+1} \in \mathbb{R}^{n \times d}$  is the message matrix.

The use of a MLP was motivated by the observation that for graph classification, MP neural nets with 1-layer perceptrons are inferior to their MLP counterparts (Xu et al. 2018a). Indeed, 1-layer perceptrons are not universal approximators

<sup>2</sup>except for words at the end/beginning of two successive sentences.

<sup>3</sup>at  $t=0$ ,  $d$  is equal to the dimensionality of the pretrained word embeddings.

<sup>4</sup>the feature of the node itself is already taken into account by our GRU-based COMBINE function (see Eq. 5).

of multiset functions. Note that like in (Xu et al. 2018a), we use a different MLP at each layer.

**Renormalization.** The rows of  $\mathbf{D}^{-1}\mathbf{A}$  sum to 1. This is equivalent to the renormalization trick of (Kipf and Welling 2016), but using only the in-degrees. That is, instead of computing a weighted sum of the incoming neighbors' feature vectors, we compute a weighted average of them. The coefficients are proportional to the strength of co-occurrence between words. One should note that by averaging, we lose the ability to distinguish between different neighborhood structures in some special cases, that is, we lose *injectivity*. Such cases include neighborhoods in which all nodes have the same representations, and neighborhoods of different sizes containing various representations in equal proportions (Xu et al. 2018a). As suggested by the results of an ablation experiment, averaging is better than summing in our application (see subsection 5.2). Note that instead of simply summing/averaging, we also tried using GAT-like attention (Veličković et al. 2017) in early experiments, without obtaining better results.

As far as our COMBINE function, we use the Gated Recurrent Unit (Cho et al. 2014; Chung et al. 2014):

$$\mathbf{H}^{t+1} = \text{GRU}(\mathbf{H}^t, \mathbf{M}^{t+1}) \quad (5)$$

Omitting biases for readability, we have:

$$\begin{aligned} \mathbf{R}^{t+1} &= \sigma(\mathbf{W}_R^{t+1}\mathbf{M}^{t+1} + \mathbf{U}_R^{t+1}\mathbf{H}^t) \\ \mathbf{Z}^{t+1} &= \sigma(\mathbf{W}_Z^{t+1}\mathbf{M}^{t+1} + \mathbf{U}_Z^{t+1}\mathbf{H}^t) \\ \tilde{\mathbf{H}}^{t+1} &= \tanh(\mathbf{W}^{t+1}\mathbf{M}^{t+1} + \mathbf{U}^{t+1}(\mathbf{R}^{t+1} \odot \mathbf{H}^t)) \\ \mathbf{H}^{t+1} &= (1 - \mathbf{Z}^{t+1}) \odot \mathbf{H}^t + \mathbf{Z}^{t+1} \odot \tilde{\mathbf{H}}^{t+1} \end{aligned} \quad (6)$$

where the  $\mathbf{W}$  and  $\mathbf{U}$  are trainable weight matrices not shared across time steps,  $\sigma(\mathbf{x}) = 1/(1 + \exp(-\mathbf{x}))$  is the sigmoid function, and  $\mathbf{R}$  and  $\mathbf{Z}$  are the parameters of the reset and update gates. The reset gate controls the amount of information from the previous time step (in  $\mathbf{H}^t$ ) that should propagate to the candidate representations,  $\tilde{\mathbf{H}}^{t+1}$ . The new representations  $\mathbf{H}^{t+1}$  are finally obtained by linearly interpolating between the previous and the candidate ones, using the coefficients returned by the update gate.

**Interpretation.** Updating node representations through a GRU should in principle allow nodes to encode a combination of local and global signals (low and high values of  $t$ , resp.), by allowing them to remember about past iterations. In addition, we also explicitly consider node representations at all iterations when reading out (see Eq. 8).

#### 3.3 Readout

After passing messages and performing updates for  $T$  iterations, we obtain a matrix  $\mathbf{H}^T \in \mathbb{R}^{n \times d}$  containing the final vertex representations. Let  $\hat{G}$  be graph  $G$  without the special document node and its adjacent edges, and matrix  $\hat{\mathbf{H}}^T \in \mathbb{R}^{(n-1) \times d}$  be the corresponding representation matrix (i.e.,  $\mathbf{H}^T$  without the row of the document node).

We use as our READOUT function the concatenation of self-attention applied to  $\hat{\mathbf{H}}^T$  with the final document node representation. More precisely, we apply a global self-attention mechanism (Lin et al. 2017) to the rows of  $\hat{\mathbf{H}}^T$ .

As shown in Eq. 7,  $\hat{\mathbf{H}}^T$  is first passed to a dense layer parameterized by matrix  $\mathbf{W}_A^T \in \mathbb{R}^{d \times d}$ . An alignment vector  $\alpha$  is then derived by comparing, via dot products, the rows of the output of the dense layer  $\mathbf{Y}^T \in \mathbb{R}^{(n-1) \times d}$  with a trainable vector  $\mathbf{v}^T \in \mathbb{R}^d$  (initialized randomly) and normalizing with a softmax. The normalized alignment coefficients are finally used to compute the attentional vector  $\mathbf{u}^T \in \mathbb{R}^d$  as a weighted sum of the final representations  $\hat{\mathbf{H}}^T$ .

$$\begin{aligned} \mathbf{Y}^T &= \tanh(\hat{\mathbf{H}}^T \mathbf{W}_A^T) \\ \alpha_i^T &= \frac{\exp(\mathbf{Y}_i^T \cdot \mathbf{v}^T)}{\sum_{j=1}^{n-1} \exp(\mathbf{Y}_j^T \cdot \mathbf{v}^T)} \\ \mathbf{u}^T &= \sum_{i=1}^{n-1} \alpha_i^T \hat{\mathbf{H}}_i^T \end{aligned} \quad (7)$$

Note that we tried with multiple context vectors, i.e., with a matrix  $\mathbf{V}^T$  instead of a vector  $\mathbf{v}^T$ , like in (Lin et al. 2017), but results were not convincing, even when adding a regularization term to the loss to favor diversity among the rows of  $\mathbf{V}^T$ .

**Master node skip connection.**  $\mathbf{h}_G^T \in \mathbb{R}^{2d}$  is obtained by concatenating  $\mathbf{u}^T$  and the final master node representation. That is, the master node vector bypasses the attention mechanism. This is equivalent to a skip or shortcut connection (He et al. 2016). The reason behind this choice is that we expect the special document node to learn a high-level summary about the document, such as its size, vocabulary, etc. (more details are given in subsection 5.2). Therefore, by making the master node bypass the attention layer, we directly inject global information about the document into its final representation.

**Multi-readout.** (Xu et al. 2018a), inspired by Jumping Knowledge Networks (Xu et al. 2018b), recommend to not only use the final representations when performing readout, but also that of the earlier steps. Indeed, as one iterates, node features capture more and more global information. However, retaining more local, intermediary information might be useful too. Thus, instead of applying the readout function only to  $t = T$ , we apply it to all time steps and concatenate the results, finally obtaining  $\mathbf{h}_G \in \mathbb{R}^{T \times 2d}$ :

$$\mathbf{h}_G = \text{CONCAT}(\text{READOUT}(\mathbf{H}^t) \mid t = 1 \dots T) \quad (8)$$

In effect, with this modification, we take into account features based on information aggregated from subtrees of different heights (from 1 to  $T$ ), corresponding to local and global features.

### 3.4 Hierarchical variants of MPAD

Through the successive MP iterations, it could be argued that MPAD implicitly captures some soft notion of the hierarchical structure of documents (words  $\rightarrow$  bigrams  $\rightarrow$  compositions of bigrams, etc.). However, it might be beneficial to explicitly capture document hierarchy. Hierarchical architectures have brought significant improvements to many NLP tasks, such as language modeling and generation (Lin et al. 2015; Li, Luong, and Jurafsky 2015), sentiment and topic classification (Tang, Qin, and Liu 2015;

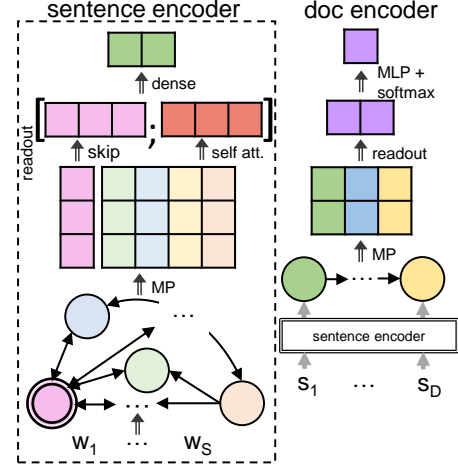


Figure 1: Illustration of MPAD-path ( $\odot$ : master node).

Yang et al. 2016), and spoken language understanding (Raheja and Tetreault 2019; Shang et al. 2019). Inspired by this line of research, we propose several hierarchical variants of MPAD, detailed in what follows. In all of them, we represent each sentence in the document as a word co-occurrence network, and obtain an embedding for it by applying MPAD as previously described.

**MPAD-sentence-att.** Here, the sentence embeddings are simply combined through self-attention.

**MPAD-clique.** In this variant, we build a complete graph where each node represents a sentence. We then feed that graph to MPAD, where the feature vectors of the nodes are initialized with the sentence embeddings previously obtained.

**MPAD-path.** This variant, shown in Fig. 1, is similar to the clique one, except that instead of a complete graph, we build a path according to the natural flow of the text. That is, two nodes are linked by a directed edge if the two sentences they represent follow each other in the document.

Note that the sentence graphs in MPAD-clique and MPAD-path do not feature a master node.

## 4 Experiments

### 4.1 Datasets

We evaluate the quality of the document embeddings learned by MPAD on 10 document classification datasets, covering the topic identification, coarse and fine sentiment analysis and opinion mining, and subjectivity detection tasks. We briefly introduce the datasets next. Their statistics are reported in Table 1.

(1) **Reuters** contains stories from the Reuters news agency. We used the ModApte split, removed documents belonging to multiple classes and considered only the 8 classes with the highest number of training examples.

(2) **BBCSport** (Greene and Cunningham 2006) contains sports news articles from the BBC Sport website.

(3) **Polarity** (Pang and Lee 2005) features positive and negative labeled snippets from Rotten Tomatoes.

Dataset	# training examples	# test examples	# classes	av. # words	max # words	voc. size	# pretrained words
Reuters	5,485	2,189	8	102.3	964	23,585	15,587
Snippets	10,060	2,280	8	18.0	50	29,257	17,142
BBCSport	737	CV	5	380.5	1,818	14,340	13,390
Polarity	10,662	CV	2	20.3	56	18,777	16,416
Subjectivity	10,000	CV	2	23.3	120	21,335	17,896
MPQA	10,606	CV	2	3.0	36	6,248	6,085
IMDB	25,000	25,000	2	254.3	2,633	141,655	104,391
TREC	5,452	500	6	10.0	37	9,593	9,125
SST-1	157,918	2,210	5	7.4	53	17,833	16,262
SST-2	77,833	1,821	2	9.5	53	17,237	15,756
Yelp2013	301,514	33,504	5	143.7	1,184	48,212	48,212

Table 1: Statistics of the datasets used in our experiments. CV indicates that cross-validation was used. # pretrained words refers to the number of words in the vocabulary having an entry in the Google News word vectors (except for Yelp2013).

(4) **Subjectivity** (Pang and Lee 2004) contains movie review snippets from Rotten Tomatoes (subjective sentences), and IMDB plot summaries (objective sentences).

(5) **MPQA** (Wiebe, Wilson, and Cardie 2005) is made of positive and negative phrases, annotated as part of the summer 2002 NRRC Workshop on Multi-Perspective Question Answering.

(6) **IMDB** (Maas et al. 2011) is a collection of highly polarized movie reviews (positive/negative).

(7) **TREC** (Li and Roth 2002) consists of questions that are classified into 6 different categories.

(8) **SST-1** (Socher et al. 2013) contains the same snippets as Polarity, split into multiple sentences and annotated with fine-grained polarity (from very negative to very positive).

(9) **SST-2** (Socher et al. 2013) is the same as SST-1 but with neutral reviews removed and snippets classified as positive or negative.

(10) **Yelp2013** (Tang, Qin, and Liu 2015) features reviews obtained from the 2013 Yelp Dataset Challenge.

## 4.2 Baselines

We evaluate MPAD against multiple state-of-the-art baseline models, including hierarchical ones, to enable fair comparison with the hierarchical MPAD variants.

**Doc2vec** (Le and Mikolov 2014) is an extension of word2vec that learns vectors for documents in a fully unsupervised manner. Document embeddings are then fed to a logistic regression classifier.

**CNN** (Kim 2014). 1D convolutional neural network where the word embeddings are used as channels (depth dimensions).

**DAN** (Iyyer et al. 2015). The Deep Averaging Network passes the unweighted average of the embeddings of the input words through multiple dense layers and a final softmax.

**Tree-LSTM** (Tai, Socher, and Manning 2015) is a generalization of the standard LSTM architecture to constituency and dependency parse trees.

**DRNN** (Irsoy and Cardie 2014). Recursive neural networks are stacked and applied to parse trees.

**LSTMN** (Cheng, Dong, and Lapata 2016) is an extension of the LSTM model where the memory cell is replaced by a memory network which stores word representations.

**C-LSTM** (Zhou et al. 2015) combines convolutional and recurrent neural networks. The region embeddings provided by a CNN are fed to a LSTM.

**SPGK** (Nikolentzos et al. 2017) also models documents as word co-occurrence networks. It computes a graph kernel that compares shortest paths extracted from the word co-occurrence networks and then relies on a SVM.

**WMD** (Kusner et al. 2015) is an application of the well-known Earth Mover’s Distance to text. A  $k$ -nearest neighbor classifier is used.

**DiSAN** (Shen et al. 2018) uses directional self-attention along with multi-dimensional attention to generate document representations.

**LSTM-GRNN** (Tang, Qin, and Liu 2015) is a hierarchical model where sentence embeddings are obtained with a CNN and a GRU-RNN is fed the sentence representations to obtain a document vector.

**HN-ATT** (Yang et al. 2016) is another hierarchical model, where the same encoder architecture (a bidirectional GRU-RNN) is used for both sentences and documents. Self-attention is applied at each level.

## 4.3 Model configuration and training

We preprocess all datasets using the code of (Kim 2014). On Yelp2013, we also replace all tokens appearing strictly less than 6 times with a special UNK token, like in (Yang et al. 2016). We then build a directed word co-occurrence network from each document, with a window of size 2.

We use two MP iterations ( $T=2$ ) for the basic MPAD, and two MP iterations at each level, for the hierarchical variants. The output of the readout goes through a dense layer before reaching the final classification layer (or the next level, at the first level of MPAD-path and MPAD-clique). We set  $d$  to 64, except on IMDB and Yelp on which  $d = 128$ , and use a two-layer MLP. The final graph representations are passed through a softmax for classification. All our dense layers (except in self-attention) use ReLU activation. We train MPAD in an end-to-end fashion by minimizing the cross-entropy loss function with the Adam optimizer (Kingma and Ba 2014) and an initial learning rate of 0.001.

To regulate potential differences in magnitude, we apply batch normalization after concatenating the feature vector of the master node with the self-attentional vector, that is,

after the skip connection (see subsection 3.3). To prevent overfitting, we use dropout (Srivastava et al. 2014) with a rate of 0.5. We select the best epoch, capped at 200, based on the validation accuracy. When cross-validation is used (see 3<sup>rd</sup> column of Table 1), we construct a validation set by randomly sampling 10% of the training set of each fold.

On all datasets except Yelp2013, we use the publicly available<sup>5</sup> 300-dimensional pre-trained Google News vectors (Mikolov et al. 2013) to initialize the node representations  $\mathbf{H}^0$ . On Yelp2013, we follow (Yang et al. 2016) and learn our own word vectors from the training and validation sets with the gensim implementation of word2vec (Řehůřek and Sojka 2010). MPAD was implemented in Python 3.6 using the PyTorch library. All experiments were run on a single machine consisting of a 3.4 GHz Intel Core i7 CPU with 16 GB of RAM and an NVidia GeForce Titan Xp GPU.

## 5 Results and ablations

### 5.1 Results

Experimental results are shown in Table 2. For the baselines, we provide the scores reported in the original papers. Furthermore, we have evaluated some of the baselines on the rest of our benchmark datasets, and we also report these scores. MPAD reaches best performance on 5 out of 10 datasets, and is close second elsewhere. Moreover, the 5 datasets on which MPAD ranks first widely differ in training set size, number of categories, and prediction task (topic, sentiment, etc.), which indicates that MPAD can perform well in different settings.

**MPAD vs. hierarchical variants.** On 9 datasets out of 10, one or more of the hierarchical variants outperform the vanilla MPAD architecture, highlighting the benefit of explicitly modeling the hierarchical nature of documents.

However, on Subjectivity, standard MPAD outperforms all hierarchical variants. On TREC, it reaches the same accuracy. We hypothesize that in some cases, using a different graph to separately encode each sentence might be worse than using one single graph to directly encode the document. Indeed, in the single document graph, some words that never appear in the same sentence can be connected through common neighbors, as was explained in subsection 3.1. So, this way, some notion of cross-sentence context is captured while learning representations of words, bigrams, etc. at each MP iteration. This creates better informed representations, resulting in a better document embedding. With the hierarchical variants, on the other hand, each sentence vector is produced in isolation, without any contextual information about the other sentences in the document. Therefore, the final sentence embeddings might be of lower quality, and as a group might also contain redundant/repeated information. When the sentence vectors are finally combined into a document representation, it is too late to take context into account.

### 5.2 Ablation studies

To understand the impact of some hyperparameters on performance, we conducted additional experiments on

the Reuters, Polarity, and IMDB datasets, with the non-hierarchical version of MPAD. Results are shown in Table 3.

**Number of MP iterations.** First, we varied the number of message passing iterations from 1 to 4. We can clearly see in Table 3 that having more iterations improves performance. We attribute this to the fact that we are reading out at each iteration from 1 to  $T$  (see Eq. 8), which enables the final graph representation to encode a mixture of low-level and high-level features. Indeed, in initial experiments involving readout at  $t=T$  only, setting  $T \geq 2$  was always decreasing performance, despite the GRU-based updates (Eq. 5)<sup>6</sup>. These results were consistent with that of (Yao, Mao, and Luo 2019) and (Kipf and Welling 2016), who both are reading out only at  $t=T$  too. We hypothesize that node features at  $T \geq 2$  are too diffuse to be entirely relied upon during readout. More precisely, initially at  $t=0$ , node representations capture information about words, at  $t=1$ , about their 1-hop neighborhood (bigrams), at  $t=2$ , about compositions of bigrams, etc. Thus, pretty quickly, node features become general and diffuse. In such cases, considering also the lower-level, more precise features of the earlier iterations when reading out may be necessary.

**Undirected edges.** On Reuters, using an undirected graph leads to better performance, while on Polarity and IMDB, it is the opposite. This can be explained by the fact that Reuters is a topic classification task, for which the presence or absence of some patterns is important, but not necessarily the order in which they appear, while Polarity and IMDB are sentiment analysis tasks. To capture sentiment, modeling word order is crucial, e.g., in detecting negation.

**No master node.** Removing the master node deteriorates performance across all datasets, clearly showing the value of having such a node. We hypothesize that since the special document node is connected to all other nodes, it is able to encode during message passing a summary of the document.

**No renormalization.** Here, we do not use the renormalization trick of (Kipf and Welling 2016) during MP (see subsection 3.2). That is, Eq. 4 becomes  $\mathbf{M}^{t+1} = \text{MLP}^{t+1}(\mathbf{A}\mathbf{H}^t)$ . In other words, instead of computing a weighted average of the incoming neighbors’ feature vectors, we compute a weighted sum of them<sup>7</sup>. Unlike the mean, which captures distributions, the sum captures structural information (Xu et al. 2018a). As shown in Table 3, using sum instead of mean decreases performance everywhere, suggesting that in our application, capturing the distribution of neighbor representations is more important than capturing their structure. We hypothesize that this is the case because statistical word co-occurrence networks tend to have similar structural properties, regardless of the topic, polarity, sentiment, etc. of the corresponding documents.

**Neighbors-only.** In this experiment, we replaced the GRU COMBINE function (see Eq. 5) with the identity function. That is, we simply have  $\mathbf{H}^{t+1} = \mathbf{M}^{t+1}$ . Since  $\mathbf{A}$  has zero diagonal, by doing so, we completely ignore the previous feature of the node itself when updating its representation.

<sup>6</sup>The GRU should in principle enable nodes to retain locality in their representations, by remembering about early iterations.

<sup>7</sup>Weights are co-occurrence counts, as before.

<sup>5</sup><https://code.google.com/archive/p/word2vec>

Model	Reut.	BBC	Pol.	Subj.	MPQA	IMDB	TREC	SST-1	SST-2	Yelp'13
doc2vec (Le and Mikolov 2014)	95.34	98.64	67.30	88.27	82.57	<b>92.5</b>	70.80	48.7	87.8	57.7
CNN (Kim 2014)	97.21	98.37	<b>81.5</b>	93.4	89.5	90.28	93.6	48.0	87.2	64.89
DAN (Iyyer et al. 2015)	94.79	94.30	80.3	92.44	88.91	89.4	89.60	47.7	86.3	61.55
Tree-LSTM (Tai, Socher, and Manning 2015)	-	-	-	-	-	-	-	51.0	88.0	-
DRNN (Irsoy and Cardie 2014)	-	-	-	-	-	-	-	49.8	86.6	-
LSTMN (Cheng, Dong, and Lapata 2016)	-	-	-	-	-	-	-	47.9	87.0	-
C-LSTM (Zhou et al. 2015)	-	-	-	-	-	-	94.6	49.2	87.8	-
SPGK (Nikolentzos et al. 2017)	96.39	94.97	77.89	91.48	85.78	OOM	90.69	OOM	OOM	OOM
WMD (Kusner et al. 2015)	96.5	98.71	66.42	86.04	83.95	OOM	73.40	OOM	OOM	OOM
DiSAN (Shen et al. 2018)	97.35	96.05	80.38	<b>94.2</b>	<b>90.1</b>	83.25	94.2	<b>51.72</b>	86.76	60.51
LSTM-GRNN (Tang, Qin, and Liu 2015)	96.16	95.52	79.98	92.38	89.08	89.98	89.40	48.09	86.38	65.1
HN-ATT (Yang et al. 2016)	97.25	96.73	80.78	92.92	89.08	90.06	90.80	49.00	86.71	<b>68.2</b>
MPAD	97.07	98.37	80.24	93.46*	90.02	91.30	<b>95.60*</b>	49.09	87.80	66.16
MPAD-sentence-att	96.89	99.32	80.44	93.02	<b>90.12*</b>	91.70	<b>95.60*</b>	49.95*	<b>88.30*</b>	66.47
MPAD-clique	<b>97.57*</b>	<b>99.72*</b>	81.17*	92.82	89.96	91.87*	95.20	48.86	87.91	66.60
MPAD-path	97.44	99.59	80.46	93.31	89.81	91.84	93.80	49.68	87.75	66.80*

Table 2: Classification accuracies. Best performance per column in **bold**, \*best MPAD variant. OOM: >16GB RAM.

MPAD variant	Reut.	Pol.	IMDB
MPAD 1MP	96.57	79.91	90.57
MPAD 2MP*	97.07	80.24	<b>91.30</b>
MPAD 3MP	97.07	80.20	91.24
MPAD 4MP	<b>97.48</b>	80.52	<b>91.30</b>
MPAD 2MP undirected	97.35	80.05	90.97
MPAD 2MP no master node	96.66	79.15	91.09
MPAD 2MP no renormalization	96.02	79.84	91.16
MPAD 2MP neighbors-only	97.12	79.22	89.50
MPAD 2MP no master node skip connection	96.93	<b>80.62</b>	91.12

Table 3: Ablation results. The  $n$  in  $n$ MP refers to the number of message passing iterations. \*vanilla model (MPAD in Table 2).

That is, the update is based entirely on its neighbors. Except on Reuters (almost no change), performance always suffers, stressing the need to take into account the root node during updates and not only its neighborhood.

**No master node skip connection.** Here, the master node does not bypass the attention mechanism and is treated as a normal node. This leads to better performance on Polarity, but slightly worse performance on Reuters and IMDB.

## 6 Related work

(Kipf and Welling 2016; Atwood and Towsley 2016; Veličković et al. 2017; Hamilton, Ying, and Leskovec 2017) conduct some node classification experiments on citation networks, where nodes are scientific papers, i.e., textual data. However, text is only used to derive node feature vectors. The external graph structure, which plays a central role in determining node labels, is completely unrelated to text.

On the other hand, (Henaff, Bruna, and LeCun 2015; Defferrard, Bresson, and Vandergheynst 2016) experiment on traditional document classification tasks. They both build  $k$ -nearest neighbor similarity graphs based on the Gaussian diffusion kernel. More precisely, (Henaff, Bruna, and LeCun 2015) build one single graph where nodes are documents and distance is computed in the BoW space. Node features are then used for classification. Closer to our work, (Deffer-

rard, Bresson, and Vandergheynst 2016) represent each document as a graph. All document graphs are derived from the same underlying structure. Only node features, corresponding to the entries of the documents' BoW vectors, vary. The underlying, shared structure is that of a  $k$ -NN graph where nodes are vocabulary terms and similarity is the cosine of the word embedding vectors. (Defferrard, Bresson, and Vandergheynst 2016) then perform graph classification. However they found performance to be lower than that of a naive Bayes classifier.

(Peng et al. 2018) use a GNN for hierarchical classification into a large taxonomy of topics. This task differs from traditional document classification. The authors represent documents as unweighted, undirected word co-occurrence networks with word embeddings as node features. They then use the *spatial* GNN of (Niepert, Ahmed, and Kutzkov 2016) to perform graph classification.

The work closest to ours is probably that of (Yao, Mao, and Luo 2019). The authors adopt the *semi-supervised node classification* approach of (Kipf and Welling 2016). They build one single undirected graph from the entire dataset, with both word and document nodes. Document-word edges are weighted by TF-IDF and word-word edges are weighted by pointwise mutual information derived from co-occurrence within a sliding window. There are no document-document edges. The GNN is trained based on the cross-entropy loss computed only for the labeled nodes, that is, the documents in the training set. When the final node representations are obtained, one can use that of the test documents to classify them and evaluate prediction performance.

There are significant differences between (Yao, Mao, and Luo 2019) and our work. First, our approach is *inductive*<sup>8</sup>, not *transductive*. Indeed, while the node classification approach of (Yao, Mao, and Luo 2019) requires all test documents at training time, our graph classification model is able to perform inference on new, never-seen documents. The downside of representing documents as separate graphs,

<sup>8</sup>Note that other GNNs used in inductive settings can be found (Hamilton, Ying, and Leskovec 2017; Veličković et al. 2017).

however, is that we lose the ability to capture corpus-level dependencies. Also, our directed graphs capture word ordering, which is ignored by (Yao, Mao, and Luo 2019). Finally, the approach of (Yao, Mao, and Luo 2019) requires computing the PMI for every word pair in the vocabulary, which may be prohibitive on datasets with very large vocabularies. On the other hand, the complexity of MPAD does not depend on vocabulary size.

MPAD is also related to the Transformer’s encoder stack (Vaswani et al. 2017). Specifically, the self-attention layer in each encoder updates the representation of each term based on the representations of all the other terms in the document, and can thus be thought of as a function performing the AGGREGATE and COMBINE steps. Stacking multiple encoders can also be thought of as performing multiple MP iterations. The main difference is that the Transformer’s self-attention graph is complete, thus ignoring word order and proximity. Also, building that graph requires constructing an adjacency matrix that may become prohibitively large with long documents.

## 7 Conclusion

We proposed an application of the message passing framework to NLP, the Message Passing Attention network for Document understanding (MPAD). Experiments show that our architecture is competitive with the state-of-the-art. By processing weighted, directed word co-occurrence networks, MPAD is sensitive to word order and word-word relationship strength. To capture the hierarchical structure of documents, we also proposed three hierarchical variants of MPAD, that bring improvements over the vanilla model.

## 8 Acknowledgments

GN is supported by the project “ESIGMA” (ANR-17-CE40-0028). We thank the NVidia corporation for the donation of a GPU as part of their GPU grant program.

## References

- [Atwood and Towsley 2016] Atwood, J., and Towsley, D. 2016. Diffusion-Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 1993–2001.
- [Battaglia et al. 2016] Battaglia, P.; Pascanu, R.; Lai, M.; Rezende, D. J.; et al. 2016. Interaction Networks for Learning about Objects, Relations and Physics. In *Advances in Neural Information Processing Systems*, 4502–4510.
- [Cheng, Dong, and Lapata 2016] Cheng, J.; Dong, L.; and Lapata, M. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 551–561.
- [Cho et al. 2014] Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1724–1734.
- [Chung et al. 2014] Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv preprint arXiv:1412.3555*.
- [Defferrard, Bresson, and Vandergheynst 2016] Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information Processing Systems*, 3844–3852.
- [Duvenaud et al. 2015] Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems*, 2224–2232.
- [Gilmer et al. 2017] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning* 1263–1272.
- [Gori, Monfardini, and Scarselli 2005] Gori, M.; Monfardini, G.; and Scarselli, F. 2005. A New Model for Learning in Graph Domains. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks*, volume 2, 729–734.
- [Greene and Cunningham 2006] Greene, D., and Cunningham, P. 2006. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, 377–384.
- [Hamilton, Ying, and Leskovec 2017] Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, 1024–1034.
- [He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 770–778.
- [Henaff, Bruna, and LeCun 2015] Henaff, M.; Bruna, J.; and LeCun, Y. 2015. Deep Convolutional Networks on Graph-Structured Data. *arXiv preprint arXiv:1506.05163*.
- [Irsoy and Cardie 2014] Irsoy, O., and Cardie, C. 2014. Deep Recursive Neural Networks for Compositionality in Language. In *Advances in Neural Information Processing Systems*, 2096–2104.
- [Iyyer et al. 2015] Iyyer, M.; Manjunatha, V.; Boyd-Graber, J.; and Daumé III, H. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1681–1691.
- [Kearnes et al. 2016] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; and Riley, P. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* 30(8):595–608.
- [Kim 2014] Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746–1751.
- [Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Kipf and Welling 2016] Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- [Kusner et al. 2015] Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From Word Embeddings to Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning*, 957–966.
- [Le and Mikolov 2014] Le, Q., and Mikolov, T. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of*

- the 31st International Conference on Machine Learning, 1188–1196.
- [Li and Roth 2002] Li, X., and Roth, D. 2002. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, 1–7.
- [Li et al. 2016] Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2016. Gated graph sequence neural networks. In *Proceedings of the 4th International Conference on Learning Representations*.
- [Li, Luong, and Jurafsky 2015] Li, J.; Luong, T.; and Jurafsky, D. 2015. A Hierarchical Neural Autoencoder for Paragraphs and Documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1106–1115.
- [Lin et al. 2015] Lin, R.; Liu, S.; Yang, M.; Li, M.; Zhou, M.; and Li, S. 2015. Hierarchical Recurrent Neural Network for Document Modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 899–907.
- [Lin et al. 2017] Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.
- [Maas et al. 2011] Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.
- [Mihalcea and Tarau 2004] Mihalcea, R., and Tarau, P. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411.
- [Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119.
- [Murphy, Weiss, and Jordan 1999] Murphy, K. P.; Weiss, Y.; and Jordan, M. I. 1999. Loopy Belief Propagation for Approximate Inference: An Empirical Study. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 467–475.
- [Niepert, Ahmed, and Kutzkov 2016] Niepert, M.; Ahmed, M.; and Kutzkov, K. 2016. Learning Convolutional Neural Networks for Graphs. In *Proceedings of the 33rd International Conference on Machine Learning*, 2014–2023.
- [Nikolentzos et al. 2017] Nikolentzos, G.; Meladianos, P.; Rousseau, F.; Stavarakas, Y.; and Vazirgiannis, M. 2017. Shortest-Path Graph Kernels for Document Similarity. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1890–1900.
- [Nikolentzos et al. 2018] Nikolentzos, G.; Meladianos, P.; Tixier, A. J.-P.; Skianis, K.; and Vazirgiannis, M. 2018. Kernel Graph Convolutional Neural Networks. In *Proceedings of the 27th International Conference on Artificial Neural Networks*, 22–32.
- [Pang and Lee 2004] Pang, B., and Lee, L. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271–278.
- [Pang and Lee 2005] Pang, B., and Lee, L. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 115–124.
- [Peng et al. 2018] Peng, H.; Li, J.; He, Y.; Liu, Y.; Bao, M.; Wang, L.; Song, Y.; and Yang, Q. 2018. Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN. In *Proceedings of the 2018 World Wide Web Conference*, 1063–1072.
- [Raheja and Tetreault 2019] Raheja, V., and Tetreault, J. 2019. Dialogue Act Classification with Context-Aware Self-Attention. *arXiv preprint arXiv:1904.02594*.
- [Řehůřek and Sojka 2010] Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.
- [Scarselli et al. 2008] Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* 20(1):61–80.
- [Shang et al. 2019] Shang, G.; Tixier, A. J.-P.; Vazirgiannis, M.; and Lorré, J.-P. 2019. Energy-based Self-attentive Learning of Abstractive Communities for Spoken Language Understanding. *arXiv preprint arXiv:1904.09491*.
- [Shen et al. 2018] Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 5446–5455.
- [Socher et al. 2013] Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.
- [Srivastava et al. 2014] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.
- [Tai, Socher, and Manning 2015] Tai, K. S.; Socher, R.; and Manning, C. D. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1556–1566.
- [Tang, Qin, and Liu 2015] Tang, D.; Qin, B.; and Liu, T. 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Proceedings of the 2015 conference on Empirical Methods in Natural Language Processing*, 1422–1432.
- [Tixier et al. 2019] Tixier, A. J.-P.; Nikolentzos, G.; Meladianos, P.; and Vazirgiannis, M. 2019. Graph Classification with 2D Convolutional Neural Networks. In *Proceedings of the 28th International Conference on Artificial Neural Networks*, 578–593.
- [Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 5998–6008.
- [Veličković et al. 2017] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- [Weisfeiler and Lehman 1968] Weisfeiler, B., and Lehman, A. A. 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Tekhnicheskaya Informatsia* 2(9):12–16.
- [Wiebe, Wilson, and Cardie 2005] Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39(2-3):165–210.
- [Xu et al. 2018a] Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018a. How Powerful are Graph Neural Networks? *arXiv preprint arXiv:1810.00826*.



- [Xu et al. 2018b] Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018b. Representation Learning on Graphs with Jumping Knowledge Networks. In *Proceedings of the 35th International Conference on Machine Learning*, 5449–5458.
- [Yang et al. 2016] Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.
- [Yao, Mao, and Luo 2019] Yao, L.; Mao, C.; and Luo, Y. 2019. Graph Convolutional Networks for Text Classification. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 7370–7377.
- [Zhou et al. 2015] Zhou, C.; Sun, C.; Liu, Z.; and Lau, F. 2015. A C-LSTM Neural Network for Text Classification. *arXiv preprint arXiv:1511.08630*.