TOWARDS CROSS-MODALITY TOPIC MODELLING VIA DEEP TOPICAL CORRELATION ANALYSIS

Jun Peng[1], *Yiyi Zhou*[⋆2], *Liujuan Cao*[1], *Xiaoshuai Sun*[3], *Jinsong Su*[4], *Rongrong Ji*[2],

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Computer Science, School of Information Science and Engineering, Xiamen University
[2]Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Cognitive Science, School of Information Science and Engineering, Xiamen University
[3]School of Computer Science and Technology, Harbin Institute of Technology
[4]Software School of Xiamen University
{pengjun, yiyizhou}@stu.xmu.edu.cn, xiaoshuaisun.hit@gmail.com, {caoliujuan, jssu, rrji}@xmu.edu.cn

ABSTRACT

The cross-modality topic detection in social media retains as an open problem mainly due to the difficulty of dealing with modality independence and modality missing. In this paper, we present a novel Deep Topical Correlation Analysis (DTCA) approach, which achieves robust and accurate topic detecting for micro-blogs and handles aforementioned challenges simultaneously. In particular, bidirectional recurrent neural networks and convolutional neural networks are used to learn deep textual and visual features, respectively. Then a Canonical Correlation Analysis based fusion scheme is proposed, which has two innovations to deal with both two problems mentioned above. We further release a large-scale cross-modal twitter dataset for topic detection. Extensive and quantitative evaluations are conducted with comparisons to several state-of-the-arts and alternative approaches on this dataset. Significant performance gains are reported to demonstrate the merits of proposed approach.

Index Terms— Topic Modelling, Deep Neural Networks, Correlation Analysis

1. INTRODUCTION

Topic detection has become a research hot-spot in the past decade, which has application prospects ranging from event tracking [1], influence estimation [2] to user preference mining [3]. Especially detection on social media platform like Twitter and Weibo. A growing number of micro-blogs recently are found to be composed of short texts, images, videos and emotions, which poses new challenges to the multi-modal topic modelling.

@NME Movies First pictures of Eddie Redmayne in J.K. Rowling's 'Harry Potter' spin-off film released

Tags: Night, city, buildings, river, dark, skyline

Fig. 1: Comparison between a typical tweet from Twitter (left) and a example of LabelME (right). It is noticeable that the textual and visual content of the LabelMe image have a strong correlation, while the ones of tweet can not be directly related.

On the one hand, it's not sure whether information from different modalities is dependent in multi-view learning, co-training and multi-modal classification [4] [5] [6], as validated in the recent work of Chen [7]. Taking left picture of Fig.1 for example, a strong independence between visual and textual information can be exhibited in a tweet. On the other hand, only a few tweets have single modal data due to the various types of tweets such as texts, images, videos, emotions and mix of them. This means it's not uncommon that a tweet has modals missing if we assume that each tweet has at least textual and visual data in our model. The above two issues have challenged existing multi-modal topic prediction approaches significantly. Because most of them have a key assumption that all modalities are completed and they are highly correlated in low semantic level [8]. Therefore, it has became an emerging task to design new schemes towards multi-modal topic detection for social media like Twitter and Weibo.

In this paper, we presented a novel approach, Deep Topical Correlation Analysis (DTCA), towards robust and accu-

This work is supported by the Nature Science Foundation of China (No.61772443, No.U1705262, and No.61572410), National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No.2017J01125 and No.2018J01106).

*Corresponding Author: Yiyi Zhou.

rate topic detection for multi-modal micro-blogs, which particularly addresses issues of the modality dependency and the modality missing. Concretely, bi-directional GRU network [9] and VGG-19 [10] are used to extract the textual and visual features, respectively. Then a joint inference and prediction model which is revised from the traditional canonical correlation analysis (CCA) scheme to tackle issues mentioned above. Notably, we introduced a novel filter-gate inspired by [11] to model the modality independence and a matrix-projection based component to handle the missing modalities. Both steps serve as gate-based controllers, they can capture the intrinsic dependency among modalities and filter out the noisy modality information, while compensating missing modality feature.

The contributions of this paper are summarized as follows: First, we addressed the two key challenges of multi-modal topic modelling on social networks which are modality independence and modality missing. Second, we proposed a novel topic detection scheme, termed DTCA, with two innovative designs, the filter gate and the matrix-projection component. Third, a multi-modal dataset was proposed for topic detection on social media.

2. RELATED WORK

To model multi-modal topics, several pioneering works are proposed in the literature, including but not limited to, Correspondence LDA (Corr-LDA) [12], Multi-modal LDA [13] and supervised LDA (sLDA) [6]. Golugula et al. [14] presented a Multi-modal Document Random Field Model (MDRF), which learns cross-modal similarities from a document corpus containing multi-modal data. As another representative work, the model Replicated Softmax [15] is originated from Restricted Boltzmann Machines (RBMs) [16] with shared parameters. To overcome such a difficulty, Over-Replicated Softmax is further proposed [17], which is a two-layer Deep Boltzmann Machine (DBM). More recently, the work in [4] is further proposed to combine the merits of two DBMs for robust topic modeling. Another noticeable work is the Neural Autogressive Distribution Estimator (NADE) [18], which is also derived from the Replicated Softmax but relies on autoregressive neural network. Finally the DocNADE [19] and SupDocNADE [5] can be considered as enhanced versions of NADE, both of which introduce a word binary tree to reduce the computation complexity.

3. DEEP TOPICAL CORRELATION ANALYSIS

We present a novel Deep Topical Correlation Analysis (DTCA) scheme with two component, filter gate and matrix projection, to deal with both issues mentioned above. Fig.2 depicts the framework of the DTCA.

3.1. Visual and Textual Channels

We use the last fully-connected layer of the VGG-19 [10] as the visual representation. Specially, given an image $\boldsymbol{I}$, we denote the extracted feature $\boldsymbol{F_I}$ as:

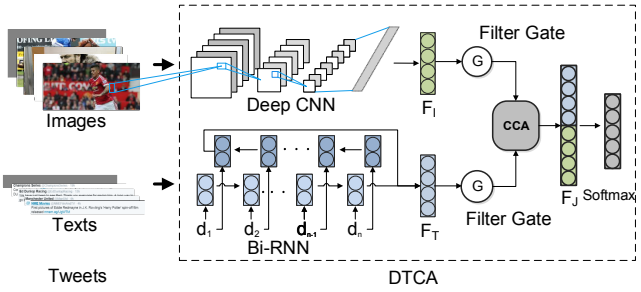

Fig. 2: The framework of our work. The pretreatment of data includes image resizing and the creation of word embeddings for representing words in dataset. In DTCA, the two subnetworks marked as "Deep CNN" and "Bi-RNN" are visual and textual channels; d_i in Bi-RNN are embedding vectors of words in the tweet; G is the filter gate component.

$$\boldsymbol{F_I} = VGG_{19}\left(I\right) \tag{1}$$

To learn the deep textual feature, we apply a bi-directional recurrent neural networks (Bi-RNNs) based on Gated Recurrent Units. The forward RNN (denoted as $\overrightarrow{RNN}$) reads the input sequence from left to right, while the backward RNN (denoted as $\overleftarrow{RNN}$) reads the sequence in reverse order. Then the text representation $\boldsymbol{F_T}$ is the concatenation of both final outputs of the forward and backward RNNs, denoted as:

$$\boldsymbol{F_T} = \left(\overrightarrow{RNN}\left(\boldsymbol{D}\right) || \overleftarrow{RNN}\left(\boldsymbol{D}\right)\right), \tag{2}$$

where $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_{n-1}, \boldsymbol{d}_n)$ are the word embeddings of a tweet, and $||$ represents the concatenation.

3.2. Filter Gate Component

As the key design of the proposed scheme, the filter gates are implemented upon the deep visual and textual channels, which targets at identifying and evaluating the modality dependency.

The proposed gate component is inspired by the recent progress in [9] [20] [21]. It serves as a denoising function over the features of different modalities. Concretely, the filter gate operation is defined as below:

$$\boldsymbol{G} = sigmoid\left(\boldsymbol{W}_g \times \boldsymbol{F} + \boldsymbol{b}_g\right), \tag{3}$$

where $\boldsymbol{F}$ can be a high-level semantic feature extracted from the visual or the textual channels. $\boldsymbol{W}_g$ and $\boldsymbol{b}_g$ are the weight and the bias of the gate. Values of elements in $\boldsymbol{G}$ range from 0 to 1. Then the filtered features $\widetilde{\boldsymbol{F}}$ can be represented as:

$$\widetilde{\boldsymbol{F}} = \boldsymbol{G} \odot \boldsymbol{F}, \tag{4}$$

where $\odot$ denotes the element-wise multiplication.

3.3. Matrix Projection Component

We further proposed the matrix projection based component to handle the modality missing. Take the case of textual data missing for instance, where the procedure is same for visual feature missing. Firstly, we uses the pre-trained visual

and textual channels to extract features of full-paired examples where $\boldsymbol{F}_I$ and $\boldsymbol{F}_T$ indicates visual and textual feature respectively. Assumed that the text is missing in a tweet which means there is not $\boldsymbol{F}_T$ for us. So we make a transformation of $\boldsymbol{F}_I$: First multiply the $\boldsymbol{F_I}$ by the projection matrix $\boldsymbol{W}_{p_I}$. Then by adding a bias term $\boldsymbol{b}_s$, we apply an element-wise activation function $\sigma = ReLU(\cdot)$. Next, we minimize the Euclidean distance between the transformed $\boldsymbol{F}_I$ and the existing textual feature $\boldsymbol{F}_T$. The objective function of this non-linear projection can be expressed as:

$$\min \left\| \boldsymbol{F}_T - \sigma \left(\boldsymbol{W}_p \cdot \boldsymbol{F}_I + \boldsymbol{b}_s \right) \right\|_2 . \tag{5}$$

When this non-linear mapping layer is fully trained, we can produce a fake text feature as a supplement for CCA-based fusion.

3.4. Correlation Analysis and Topic Modeling

Towards the efficient fusion of two modalities, we extend the traditional Canonical Correlation Analysis to our deep multimodal networks. Concretely, given an image feature $\widetilde{\mathbf{F}}_{\mathbf{I}} \in \mathbb{R}^{n \times q}$ and a text feature $\widetilde{\mathbf{F}}_{\mathbf{T}} \in \mathbb{R}^{n \times p}$, we first covert both matrices into the same feature dimension:

$$\begin{aligned} \hat{\mathbf{F}}_I &= \sigma \left(\widetilde{\mathbf{F}}_I \boldsymbol{W}_{MI} + \boldsymbol{b}_I \right) \\ \hat{\mathbf{F}}_T &= \sigma \left(\widetilde{\mathbf{F}}_T \boldsymbol{W}_{MT} + \boldsymbol{b}_T \right), \end{aligned} \tag{6}$$

where σ can be any activation function. $\boldsymbol{W}_{MI} \in \mathbb{R}^{q \times q}$, $\boldsymbol{W}_{MT} \in \mathbb{R}^{p \times q}$ and $\boldsymbol{b}_I \in \mathbb{R}^q, \boldsymbol{b}_T \in \mathbb{R}^q$ are weights and biases.

Then, we aim to find a pair of canonical variables $\mathbf{u} = \hat{\mathbf{F}}_I \cdot \boldsymbol{a}$, and $\mathbf{v} = \hat{\mathbf{F}}_T \cdot \boldsymbol{b}$, where $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{1 \times q}$ are canonical vector that are used to change the direction of $\hat{\mathbf{F}}_I$ and $\hat{\mathbf{F}}_T$ in the semantic space. The correlation between two types of features can be modelled as:

$$cor\left(\mathbf{u}, \mathbf{v}\right) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \|\mathbf{v}\|}, \tag{7}$$

The corresponding optimization problem can be formulated as:

$$\boldsymbol{o}_c = \boldsymbol{a}^T \hat{\mathbf{F}}_I^T \hat{\mathbf{F}}_T \boldsymbol{b}. \tag{8}$$

When the topic correlation of two modalities are maximized, $\widetilde{\mathbf{F}}_I$ or $\widetilde{\mathbf{F}}_T$ can be used as the joint representation. In terms of topic classification, we use the cross entropy as the cost. Then the objective function of the overall framework is:

$$\boldsymbol{o} = \boldsymbol{o}_e - \beta \cdot \boldsymbol{o}_c + \alpha \left(\|\boldsymbol{W}_{MI}\|_2 + \|\boldsymbol{W}_{MT}\|_2 \right) \tag{9}$$

Notably, our DTCA might be easily extend to scenarios where more types of modalities are involved like audio and videos.

4. EXPERIMENT

To validate the performance of our model, we create a large-scale multi-modal dataset.[1] Here we term it as Topic Modelling Twitter, TM-Twitter. We compare our model with state-of-the-arts, and the experiment results validate merits of our approach.

Table 1: Details of the Twitter Dataset

Topic	Twitter Accounts	Amount
Sport	@espn, @BBCSport, @Sport_EN, @TwitterSports	6000
Health	@cnnhealth, @bbchealth, @WSJhealth	6000
Film	@IMDb, @Film4, @BBCFilms, @TelegraphFilm	6000
Music	@NME, @billboard, @BBC6Music, @guardianmusic	6000
Politics	@HuffPostPol, @BBCPolitics, @CNNPolitics, @nprpolitics	6000

4.1. TM-Twitter

Our dataset consists of five categories which are "Sports", "Music", "Film", "Health" and "Politics". We take samples containing both texts and images from public news accounts, such as "BBCSport" and "CNNPolitics", and tag them based on their account categories. Each account only publishes the content of a particular topic, not multiple topics. In addition, we manually check the collected data and filter out inappropriate data, for example, the image contains only characters. Finally, we have 30,000 examples, each with 6,000 examples. For each class, we have 4000 training examples, 1000 verification examples, and 1000 test cases. Details are shown in Table.1.

4.2. Compared Methods

The state-of-the-arts used in our experiment are as follows: Multimodal Deep Boltzmann Machines (M-DBM) [4], Supervised Document Neural Autoregressive Distribution Estimator(SupDocNADE) [5], Supervised LDA (S-LDA) [6], Document Neural Autoregressive Distribution Estimator (DocNADE) [19].

In order to comprehensively evaluate the merits of our model, we also propose several combinatorial approaches:

- VGG19-ONLY: Only the visual channel is used.
- Bi-RNN-ONLY: Only textual channel is used.
- SG-ONLY: Use the Skip-Gram algorithm with only text data.
- VGG19+Bi-RNN: Both visual and textual channels are used, and the outputs of two channels are concatenated as the model output.

4.3. Experiment Setup

DTCA. In the preprocessing step, images are resized to $224 \times 224 \times 3$. Punctuations, numbers and prepositions in all tweets are removed, and then embedding vectors of words left are created using the Skip-Gram model. The dimension of word embeddings is 200. In terms of the visual channel setting, we initialize all layers in VGG19 with weights pre-trained on ImageNet, and then the weights of all convolutional layers are fixed. Additional dropout layers with a ratio of 0.8 are added after each forward layers. In the textual channel, the dimension of two direction channels is 256, and the one of the final output is 512. In the setting of the CCA based scheme, the dimension of the final joint representation is 1000, and

[1] Data of this dataset is collected from Twitter by using the public Twitter API.

Table 2: Comparisons of Different Methods

Method	Accuracy
M-DBM [4]	56.40%
SupDocNADE [5]	58.05%
S-LDA [6]	43.51%
DocNADE [19]	35.24%
SG-ONLY	51.21%
Bi-RNN-ONLY	73.56%
VGG19-ONLY	78.33%
VGG19+Bi-RNN	79.48%
DTCA(ours)	82.89%

a dropout layer is also added before the Softmax layer. The learning rate of DTCA is 10^{-4}, and the number of training epoch is 100. The value of α and β in Eq.9 are selected via cross-validations.

4.4. Results

The experimental results are shown in Table.2. Our proposed DTCA achieves the best performance among all approaches. Compared with state-of-the-arts, these results prove that the feature extraction strategy of DTCA will be more applicable to multi-modality data in social media. Most of alternative approaches proposed in this paper also outperform state-of-the-arts, which implies the important role of deep network based structure towards the feature extraction of high-semantic information. Bi-RNN-ONLY greatly outperforms SG-ONLY, which indicates that the bidirectional RNN used in our work is capable of learning the high-semantic feature of a tweets' text well. Notably, the performance of VGG19-ONLY is better than that of Bi-RNN-ONLY, which subverts the common assumption that the text content is more discriminative than the visual content. This result may be due to the informal and short format of tweets and the limit number of training data.

Fig.3 (left) reflects the relation between the accuracy and correlation rate. From the figure we can see that as the correlation increases, the accuracy increases and reaches the top with a correlation value between 0.7 and 0.8. After that, it starts to decrease. Fig.3 (right) reflects the relation between the accuracy and the dimensions of the joint representations. As shown in this figure, joint representations with too small dimension size are unable to capture information of two modalities well. When the dimension is too large, the joint representations contain more noises and result in a decreased classification performance.

Examples of correct or incorrect classification results are shown in Fig.4. From the correct results, we can find that DTCA has been able to learn high-level semantic features between two modalities. The incorrect examples imply the ambiguity problem of Twitter data. Take the incorrect instance of "Film" row as an example, the visual content is very like a screen-shot of a film and the textual content did not present an obvious tendency. Hence, the labelling of such kind of examples heavily depends on individual recognitions and personal knowledge backgrounds. Nevertheless, it is noticeable to

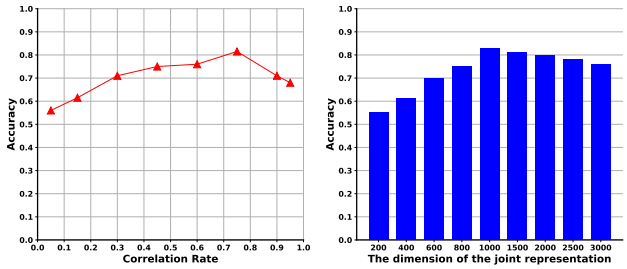

Fig. 3: Sensitivity w.r.t. to correlation rates and the feature dimensions.

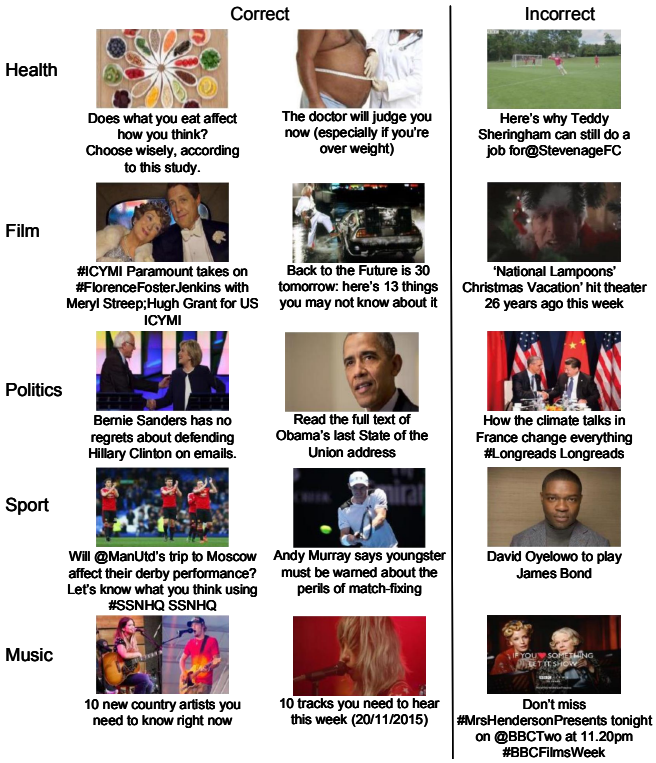

Fig. 4: Examples of classification results.

find that DTCA is capable to detect inaccurately labelling examples, for instance, the incorrect example of "Politics" row which are labelled as "Music" topic for it is from the 'RollingStone' account.

5. CONCLUSION

In this paper, we proposed a Deep Topical Correlation Analysis approach towards topic modelling in microblog-based social platforms like Twitter and Weibo. Two innovative designs, the filter gate and the matrix-projection based component, are proposed in our scheme to deal with the two key challenges namely, the modality dependency and the modality missing. Meanwhile, we propose a large-scale datasets for multi-modal topic detections on social media. On this dataset, superior performances are obtained, which confirms the merits of our scheme.

6. REFERENCES

[1] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and Jie Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

[2] Chiara Ravazzi, Roberto Tempo, and Fabrizio Dabbene, "Influence estimation in sparse social networks," pp. 775–780, 2017.

[3] Yue Ma, Guoqing Chen, and Qiang Wei, "Finding users preferences from large-scale online reviews for personalized recommendation," *Electronic Commerce Research*, vol. 17, no. 1, pp. 3–29, 2017.

[4] Nitish Srivastava and Ruslan Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Journal of Machine Learning Research*, vol. 15, no. 8, pp. 1967 – 2006, 2014.

[5] Yin Zheng, Yu Jin Zhang, and Hugo Larochelle, "Topic modeling of multimodal data: An autoregressive approach," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 1370–1377.

[6] Chong Wang, D. Blei, and Fei Fei Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1903–1910.

[7] Fuhai Chen, Yue Gao, Donglin Cao, and Rongrong Ji, "Multimodal hypergraph learning for microblog sentiment prediction," in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, 2015, pp. 1–6.

[8] St, "In between modes: Language and image in printed media," .

[9] Kyunghyun Cho and Van Merri, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," .

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[11] Duyu Tang, Bing Qin, and Ting Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1422–1432.

[12] David M Blei and Michael I Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134.

[13] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Multimodal semi-supervised learning for image classification," in *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society, 2010, pp. 902–909.

[14] A Golugula, G Lee, S. R. Master, M. D. Feldman, J. E. Tomaszewski, and A Madabhushi, "Supervised regularized canonical correlation analysis: Integrating histologic and proteomic data for predicting biochemical failures," in *Conference: International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference*, 2011, pp. 6434–6437.

[15] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Replicated softmax: an undirected topic model," in *Advances in neural information processing systems*, 2009, pp. 1607–1614.

[16] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 791–798.

[17] Ruslan Salakhutdinov and Geoffrey E Hinton, "Deep boltzmann machines," in *International conference on artificial intelligence and statistics*, 2009, pp. 448–455.

[18] Hugo Larochelle and Iain Murray, "The neural autoregressive distribution estimator.," .

[19] Hugo Larochelle and Stanislas Lauly, "A neural autoregressive topic model," in *Advances in Neural Information Processing Systems*, 2012, pp. 2708–2716.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," pp. 1096–1103, 2008.