Image-based Text Classification using 2D Convolutional Neural Networks

Erinç Merdivan[*¶], Anastasios Vafeiadis[†], Dimitrios Kalatzis[†], Sten Hanke[†], Johannes Kropf[*], Konstantinos Votis[†], Dimitrios Giakoumis[†], Dimitrios Tzovaras[†], Liming Chen[‡], Raouf Hamzaoui[‡] and Matthieu Geist[§]
[*]*Austrian Institute of Technology,GmbH - Wiener Neustadt,Austria*
Email: {erinc.merdivan, sten.hanke, johannes.kropf}@ait.ac.at
[†]*Information Technologies Institute - Center of Research & Technology Hellas- Thessaloniki, Greece*
Email: {anasvaf, dkal, kvotis, dgiakoum, tzovaras}@iti.gr
[‡]*Faculty of Computing, Engineering and Media - De Montfort University - Leicester, UK*
Email: {liming.chen, rhamzaoui}@dmu.ac.uk
[§]*Université de Lorraine, CNRS, LIEC, F-57000 - Metz, France (now at Google Brain)*
Email: matthieu.geist@univ-lorraine.fr
[¶]*CentraleSupélec, Université de Lorraine, CNRS, LORIA, F-57000 - Metz, France*

***Abstract*—We propose a new approach to text classification in which we consider the input text as an image and apply 2D Convolutional Neural Networks to learn the local and global semantics of the sentences from the variations of the visual patterns of words. Our approach demonstrates that it is possible to get semantically meaningful features from images with text without using optical character recognition and sequential processing pipelines, techniques that traditional natural language processing algorithms require. To validate our approach, we present results for two applications: text classification and dialog modeling. Using a 2D Convolutional Neural Network, we were able to outperform the state-of-art accuracy results for a Chinese text classification task and achieved promising results for seven English text classification tasks. Furthermore, our approach outperformed the memory networks without match types when using out of vocabulary entities from Task 4 of the bAbI dialog dataset.**

1. Introduction

Recent advances in Natural Language Processing (NLP) make heavy use of neural network models. Solutions for tasks such as semantic tagging [1], text classification [2] and sentiment analysis [3] rely on either Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN) variants. In the latter case, the vast majority of the proposed models are based on character-level CNNs applied on one-hot vectors of text or 1D CNNs [4]. Although the results are promising, having either surpassed or equaled the previous state of the art, there are a few issues regarding the proposed models, which are all related to the fundamental inductive bias underlying these models' architectural design. Whether working at the word- or character-level, language processing with most neural network models almost always translates to sequential processing of a string of abstract discrete symbols.

CNNs based on 1D or character convolutions constitute the vast majority of CNN models used in language processing. These networks are fast if the dictionary size is small. However, for some languages, the one-hot encoding vector dimension for input sequences can be very large (e.g., over 3000 for Chinese characters). Furthermore, and specifically for RNN variants, training for long input sequences is difficult due to the well-known problem of vanishing gradients. While architectures like Long Short-Term Memory (LSTM) [5] and Gated Recurrent Units (GRU) [6] were specifically designed to tackle this problem, stable training on long sequences remains an elusive goal, with recent works devising yet more ways to improve performance in recurrent models [7], [8], [9]. Moreover, many state of the art recurrent models rely on the attention mechanism to improve performance [10], [11], [12], which places an additional computational burden on the overall method.

To tackle the above problems, we use CNNs to process the entire text at once as an image. In other words, we convert our textual datasets into images of the relevant documents and apply our model on raw pixel values. This allows us to sensibly apply 2D convolutional layers on text, taking advantage of advances in neural network models designed for and targeting computer vision problems. Doing so, allows us to bypass the issues stated earlier relating to the use of 1D character-level CNNs and RNNs, since now the processing of documents relies on parallel extraction of visual features of many lines (depending on filter size) of text. Regarding the vanishing gradient problem, we can take advantage of recent CNN architectural advances [13], [14], [15], which specifically aim to improve its effects. In terms of linguistics, our approach is based on the distributional hypothesis [16], where our model produces compositional hierarchies of document semantics by way of its hierarchical architecture. Beyond providing an alternative computational method to deal with the problems described above, our approach is also motivated by findings in neuroscience,

cognitive science and the medical sciences where the link between visual perception and recognition of words and semantic processing of language has long been established [17], [18]. Our approach is robust to textual anomalies, such as spelling mistakes, unconventional use of punctuation (e.g., multiple exclamation marks), etc. which factors in during feature extraction. As a result, not only is the need of laborious text preprocessing removed, but the derived models are able to capture the semantic significance of the occurrence of such phenomena (e.g., multiple exclamation marks to denote emphasis), which proves to be especially helpful in tasks such as text classification and/or sentiment analysis. Moreover, our approach can work with any text (latin and non-latin), text font, misspellings and punctuation. Furthermore, it can be extended to handwriting, background colors and table formatted text naturally. It also removes the need of pre-processing real-world documents (and thus the need for optical character recognition, spell check, stemming, and character encoding correction).

Our approach is based on the hypothesis that more semantic information can be extracted from features derived from the visual processing of text than by processing strings of abstract discrete symbols. We test this hypothesis on NLP tasks and show that a solid capture of text semantics leads to better model performance. Our contributions are summarized as follows:

- a proof of concept that text classification can be achieved over an image of the text;
- a proof of concept that basic dialogue modeling (restaurant booking), in an information retrieval setting, can be completed using only image processing methods;

The remainder of the paper is organized as follows: Section 2 positions our approach compared to related work, Section 3 introduces the proposed method, Section 4 presents the experimental results and Section 5 draws the conclusions.

2. Related Work

The use of convolutional neural networks for natural language processing has attracted increasing attention in recent years. For sentence classification, Kim [19] used a simple CNN architecture consisting of one convolutional layer with multiple filters of different sizes, followed by max-pooling. The feature maps produced are then fed to a softmax layer for classification. Despite its simplicity, this architecture exhibited good performance. Sentence modeling was further explored by Blunsom et al. [20] who used an extended application, which they call Dynamic Convolutional Neural Network (DCNN) to deal with various input lengths and short- and long-term linguistic dependencies. Wang et al. [21] perform clustering in an embedding space to derive semantic features which they then feed to a CNN with a convolutional layer, followed by k-max pooling and a softmax layer for classification.

Character-level (as opposed to word- or sentence-level) feature extraction was investigated by Zhang et al. [22] who used a standard deep convolutional architecture for text classification. Conneau et al. [23] showed that using very deep convolutional architecture improves results over standard deep convolutional networks on text classification tasks. Dos Santos and Gatti [24] carried out sentiment analysis on sentences taken from text corpora, using a CNN architecture which derives input representations that are hierarchically built from the character to the sentence level. Johnson and Zhang [25] used a CNN for text categorization. Their method does not rely on pre-trained word embeddings, but rather computes convolutions directly on high-dimensional text data represented by one-hot vectors. An architectural variation was also proposed for adapting a bag-of-words model in the convolutional layers. Johnson and Zhang [26] used CNNs for sentiment and topic classification in a semi-supervised framework, where they retained the representations derived by a CNN over text regions, and which they then integrated into the supervised CNN classifier. Ruder et al. [27] employed a novel architecture combining character- and word-level channels to determine an unseen text's author among a large number of potential candidates, a task they called large-scale authorship attribution. Bjerva et al. [28] introduced a semantic tagging method, which combines stacked neural network models and a residual bypass function. The stacked neural networks consist of a vanilla CNN or a ResNet [14] in the lower level for character-/word-level feature extraction and a bidirectional Gated Recurrent Unit (GRU) in the higher level. The residual bypass function preserves the saliency of lower-level features that could be potentially lost in the processing chain of intermediate layers.

Dialog managers can be trained either as generative models or as discriminative models to differentiate good replies in Next Utterance Classification (NUC) [29]. In generative models [30], [31], [32], dialog managers are trained to produce replies for a given dialogue history. In NUC setting, a dialogue manager needs to choose the correct response from a set of candidate replies as Memory Networks (MemNets) [33], [34] in Facebook bAbI dataset [35].

While all the aforementioned works exploited CNNs for NLP tasks, they all used text data as input, either pre-trained word embeddings or simply one-hot vector representations.

3. Method

In our approach, we treat text classification as a problem which concerns the learning of context-dependent semantic conventions of language use in a given corpus of text. We treat this complex problem as an image processing problem, where the model processes an image with the text body (Figure 2), learning both the local (word- and sentence-level) and the global semantics of the corpus. In this way, the domain or context dependent meaning of sentences is implicitly contained in the variations of the visual patterns given by the distribution of words in sentences. As such, the

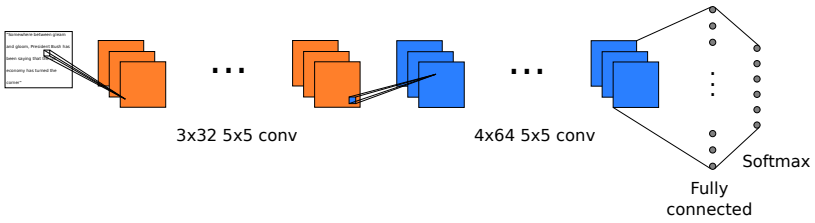

Figure 1. Proposed model: 3 convolutional layers consisting of 32 5x5 filters each, are followed by 4 convolutional layers consisting of 64 5x5 filters each. A linear fully connected layer and a classification output layer complete the model.

problem is that the model needs to observe as many variations of in-domain text as possible to be able to generalize adequately. This process is similar to the analytical method of learning to read [36], where the global meaning of a body of text is acquired first and learning of the text's meaning moves to hierarchically lower linguistic units. In our case, this translates to capturing the structure and context of the whole corpus first, then the sentences, and finally the words that constitute these sentences.

3.1. Models

For the tasks of (English and Chinese) text classification we used a vanilla CNN and also the Xception architecture [15] to check whether better vision deep networks can increase performance.

The vanilla CNN consists of seven convolutional layers, a fully connected layer and an output layer containing as many units as classes (e.g., for a classification problem with four classes, the output layer would contain four units). All filters in the convolutional layers are 5x5 with stride 2. The first three layers use 32 filters, while the rest use 64 filters. The fully connected layer consists of 128 units. All units in all layers use the rectifier function, apart from the output layer, which uses a softmax output. Figure 1 shows the architecture of the model.

For the task of dialog modeling we used version 4 of the recently proposed deep Inception network (Inception-V4) [37]. Our choice was motivated by the fact that the vanilla CNN model was too simple to effectively model the dialog structure, as well as its pragmatics (i.e., the use of language in discourse within the context of a given domain), a problem which Inception-V4 seems to have tackled, at least to a certain extent. We selected the Inception-V4 against the Xception because we wanted to experiment with different advanced architectures for similar tasks.

3.2. Data Augmentation

Data augmentation has been shown to be essential for training robust models [23], [38]. For image recognition, augmentation is applied using simple transformations such as shifting the width and the height of images by a certain percentage, scaling, or randomly extracting sub-windows from a sample of images [39].

For the task of English and Chinese text classification, we used the *ImageDataGenerator* function provided by Keras [40]. The input image was shifted in width and height by 20%, rotated by 15 degrees and flipped horizontally, using a batch size of 50. For the task of dialogue modeling, we applied the same augmentation techniques and random character flipping. Character flip and in particular changing the rating of a restaurant improved the per-response and per-dialog accuracy, especially for difficult sentences, such as booking a 4 star restaurant.

4. Results

To validate our approach, we ran experiments for two separate tasks: text classification and dialog modeling, using a single NVIDIA GTX 1080 Ti GPU.

4.1. Text classification

In this task we trained our model on an array of datasets which contained text related to news (AG's News and Sogou's News), structured ontologies on Wikipedia (DBPedia), reviews (Yelp and Amazon) and question answering (Yahoo! answers). Details about the datasets can be found in [22]. For this task, Zhang et al. [22] tested CNNs that use 1D convolutions in the task of text classification, which may more broadly include natural language processing, as well as sentiment analysis. While the model in [22] uses text as input vectors, our proposed method uses image data of text. In other words, whereas Zhang et al. [22] use one-hot vector representations of words or word embeddings, we use binarized pixel values of grayscale images of text corpora.

Table 1 shows our method's held-out accuracy in the task of Latin and Sogou News in Chinese text classification for each of the datasets. All baselines are derived from Table 4 of Zhang et al. [22] and Conneau et al. [23]. We denote the vanilla CNN by *TI-CNN* (Text-to-Image Convolutional Neural Networks). The column *Worst-Best Performance* shows the worst and best held-out accuracy achieved by the baseline models. Our approach achieved comparable results to most of the best performing baselines. The Amazon datasets were large and we did not have enough computational resources to achieve comparable results to the state-of-art with Xception.

Table 4.1 shows human generated text (not included in the training set) used for testing. For these examples,

TABLE 1. RESULTS OF LATIN AND CHINESE TEXT CLASSIFICATION IN TERMS OF HELD-OUT ACCURACY. WORST-BEST PERFORMANCE REPORTS THE RESULTS OF THE WORST AND BEST PERFORMING BASELINES FROM TABLE 4 OF ZHANG ET AL. [22] AND CONNEAU ET AL. [23]. RESULTS REPORTED FOR *TI-CNN* WERE OBTAINED IN 10 EPOCHS

Dataset	Worst-best Performance (%)	TI-CNN (%)	Xception (%)	Number of Classes
AG's News	83.1-92.3	80.0	91.8	4
Sogou News (Pinyin)	89.2-97.2	90.2	94.6	5
Sogou News (Chinese)	93.1-94.5	-	**98.0**	5
DBPedia	91.4-98.7	91.7	94.5	14
Yelp Review Polarity	87.3-95.7	90.3	92.8	2
Yelp Review Full	52.6-64.8	55.1	55.7	5
Yahoo! Answers	61.6-73.4	57.6	73.0	10
Amazon Review Full	44.1-63	50.2	57.9	5
Amazon Review Polarity	81.6-95.7	88.6	94.0	2

Sample No	Text Sample	Positivity Score	Sample No	Text Sample	Positivity Score
1	this product is mediocre	0.60	5	I love this product it is great	0.99
2	this product is excelent	0.91	6	I like this product it is ok	0.78
3	this product is excellent	0.96	7	I don't know	0.56
4	this product is excellent!!!	0.98	8	as;kdna;sdn nokorgmnsd kasdn;laknsdnaf	0.51

the table shows predictions after the model was trained on the *Amazon Review Polarity* dataset [41], which contains reviews of products in various product categories. The dataset is used for binary (positive/negative) sentiment classification of text and the metric *(positivity score)* is the probability of the positive class. The model was able to discriminate between words expressing different degrees of the same sentiment (e.g., samples 1,6 compared to samples 2-5). Sample 2 (compared to samples 3-4) illustrates our method's robustness to anomalies like spelling mistakes. In a traditional NLP setting the misspelled word would have a different representation from the respective correctly-spelled word. Unless the model was trained on data that contained many of these anomalies, or engineered by a human, it would not necessarily correlate the misspelled word with the sentiment it expressed. In our model the misspellings are handled naturally. We note that while this can be alleviated by preprocessing procedures or character-level models, these require more pre-processing or human intervention than our method.

As discussed before, the model builds these visual representations in a bottom-up fashion, creating a semantic hierarchy which is derived from language use within the context of the corpus domain. Sample 4 shows another interesting characteristic of our model which is capturing the effect of punctuation (exclamation marks) even if used informally. The exclamation marks used in sample 4 generated the highest prediction score for positive sentiment among all variations of the same phrase (samples 2-4). Samples 5 and 6 have a similar structure but the different choice of words to describe positive sentiment affects the prediction score. This also exhibits the model's capacity to build meaningful hierarchical representations, as it has learned to discriminate between the small nuances (e.g., choice of words) encountered in (visually and semantically) similar textual structures (sentences). Interestingly, an input which expresses a "neutral" sentiment, such as sample 7, has an analogous prediction score (0.56) that is closer to random guessing in a model that was trained in binary sentiment

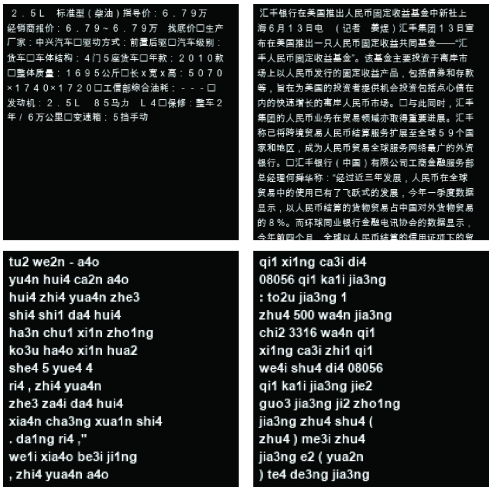

Figure 2. Top: Sogou News dataset with Chinese characters. Bottom: Sogou News dataset with pinyin

prediction, which is reasonable behavior. The model is also robust to nonsensical text such as sample 8. Finally, we applied the Xception architecture to the Sogou News dataset, using the original Chinese characters (Figure 2). Huang and Wang [4] used 1D CNNs for text classification with Chinese characters and showed that the accuracy recognition was higher than the traditional conversion to the pinyin romanization system. We extended this work by using the Xception architecture in the 2D image to achieve almost the same result (Table 1). This proves that regardless of how many Chinese words we fit in a 300x300 or a 200x200 image, our approach outperformed the NLP sequential CNNs. Furthermore, the performance improved when using the Chinese characters instead of the pinyin.

4.2. Dialog modeling

For the dialog modeling task, we tested our Inception-V4-based agent in task 4 of the bAbI dialog dataset [35],

since it requires knowledge base information when choosing the replies to the user (e.g., address, phone number). The bAbI dialog dataset consists of 1000 training, 1000 validation and 1000 test dialogs in the restaurant booking domain. Each dialog is divided in four different tasks. Here we focus on task 4, where the dialog agent should be able to read entries about restaurants from a relevant knowledge base and provide the user the requested information, such as restaurant address or phone number. We note that restaurant phone numbers and addresses have been delexicalized and replaced by tokens representing this information. We chose to focus on this task to demonstrate the increased effectiveness of visual processing of dialog as opposed to purely linguistic processing, due to the high number of different lexical tokens. In our approach the agent needs to correlate the visual pattern of a knowledge base entry to the relevant request. While in principle this should be easy to achieve using artificial delexicalized tokens, as in this benchmark task, it would be far more difficult to do so in the real world, with non-standard sequences of words (such as restaurant names, addresses etc). However, given the results of the text classification tasks, we hypothesize that given enough data, our visual approach can create semantic models that encapsulate such correlations.

As in text classification, we trained the model with images of dialog text taken from the bAbI corpus. So the agent learns the expected user utterances and their corresponding responses on the system side by processing images of in-domain dialog text. The agent learned visual representations of text meaning and structure both at word-level (implicitly, through the optimization process) and utterance-level (explicitly, through labeling of correct and incorrect responses given a user utterance).

TABLE 2. FACEBOOK BABI DIALOG TASK 4

Metrics	Inception-V4 (%)	Memory Networks w/o Match Type (%)
Per-response Accuracy	63.3	59.5
Per-dialog Accuracy	11.4	3.0

Table 2 shows the Inception-V4 performance against the MemNets used in [35]. The table shows that, our approach is competitive with MemNets when the latter does not use match types. Bordes et al. [35] introduced match types to make their model rely on type information, rather than exact match of word embeddings corresponding to words that frequently appear as containing out of vocabulary (OOV) entities (restaurant name, phone number, etc). This is because it is hard to distinguish between similar embeddings in a low-dimensional embedding space (e.g., phone numbers) as they lead to full scores in all metrics. In real life, match types would require a lexical database to identify every word type which is not realistic.

5. Conclusion

We presented a proof of concept that natural language processing can be based on visual features of text. For non-dialog text, images of text as input to CNN models can build hierarchical semantic representations which let them detect various subtleties in language use, irrespective of the language of the input data. For dialog text, we showed that CNN models learn both the structure of discourse and the implied dialog pragmatics implicitly contained within the training data. Although our model is trained in an NUC setting, it could be expanded as a generative model by using an image-based encoder for dialogue history and a language-based model for decoding. Crucially, unlike traditional NLP applications, our approach does not require any preprocessing of natural language data, such as tokenization, optical character recognition, stemming, or spell checking. Our method can work using different computer fonts, background colors and can be expanded to human handwriting. It can perform NLP tasks on real-word documents that include tables, bold, underlined and colored text, where traditional NLP methods, as well as language agnostic models (1D CNN) fail.

Our work is a first step towards expanding the methods for natural language processing, exploiting recent advances in image recognition and computer vision.of this approach are promising for a wide range of NLP tasks, such as text classification, sentiment analysis, dialog modeling and natural language processing.Future work will study the effect of pre-trained models on non-dialog corpora with regard to modeling performance, incorporation within generative frameworks for text generation for tasks like text summarization or performance of more complex networks (such as recurrent CNNs). For the task of text classification, the recognition accuracy of the Xception will increase, since the reported results are achieved without any fine-tuning for the specific datasets. As computer vision deep learning models continue to improve, we expect our results for the NLP task to follow suit.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157, project ACROSSING.

References

[1] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[2] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[3] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[4] W. Huang, "Character-level convolutional network for text classification applied to chinese corpus," Ph.D. dissertation, University College London, 2016.

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

[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[7] M. Seo, S. Min, A. Farhadi, and H. Hajishirzi, "Neural speed reading via skim-RNN," in *International Conference on Learning Representations*, 2018.

[8] T. Trinh, A. Dai, T. Luong, and Q. Le, "Learning longer-term dependencies in RNNs with auxiliary losses," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmssan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 4965–4974.

[9] A. W. Yu, H. Lee, and Q. Le, "Learning to skim text," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 1880–1890.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.

[11] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 1412–1421.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[16] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[17] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. D. Friederici, "Music, language and meaning: brain signatures of semantic processing," *Nature neuroscience*, vol. 7, no. 3, p. 302, 2004.

[18] P. G. Simos, L. F. Basile, and A. C. Papanicolaou, "Source localization of the n400 response in a sentence-reading paradigm using evoked magnetic fields and magnetic resonance imaging," *Brain research*, vol. 762, no. 1-2, pp. 29–39, 1997.

[19] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[20] P. Blunsom, E. Grefenstette, and N. Kalchbrenner, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

[21] P. Wang, J. Xu, B. Xu, C. Liu, H. Zhang, F. Wang, and H. Hao, "Semantic clustering and convolutional neural network for short text categorization," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, 2015, pp. 352–357.

[22] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[23] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, 2017, pp. 1107–1116.

[24] C. N. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts." in *COLING*, 2014, pp. 69–78.

[25] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 103–112.

[26] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *Advances in neural information processing systems*, 2015, pp. 919–927.

[27] S. Ruder, P. Ghaffari, and J. G. Breslin, "Character-level and multi-channel convolutional neural networks for large-scale authorship attribution," *arXiv preprint arXiv:1609.06686*, 2016.

[28] J. Bjerva, B. Plank, and J. Bos, "Semantic tagging with deep residual networks," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 3531–3541.

[29] R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "On the evaluation of dialogue systems with next utterance classification," *Proceedings of the 2016 SIGDIAL*, 2016.

[30] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[31] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[32] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," *arXiv preprint arXiv:1701.06547*, 2017.

[33] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," in *Advances in neural information processing systems*, 2015, pp. 2440–2448.

[34] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv preprint arXiv:1410.3916*, 2014.

[35] A. Bordes and J. Weston, "Learning end-to-end goal-oriented dialog," in *International Conference on Learning Representations*, 2017.

[36] S. Cèbe and R. Goigoux, *Apprendre à lire à l'école: Tout ce qu'il faut savoir pour accompagner l'enfant*. Retz, 2011.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[38] A. Copestake, "Augmented and alternative nlp techniques for augmentative and alternative communication," *Natural Language Processing for Communication Aids*, 1997.

[39] Í. de Pontes Oliveira, J. L. P. Medeiros, and V. F. de Sousa, "A data augmentation methodology to improve age estimation using convolutional neural networks," in *Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on*. IEEE, 2016, pp. 88–95.

[40] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[41] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 165–172.