Computer Vision Invasion into Natural Language Processing in Deep Learning

Seung Hwan Lee

December 2017

Abstract

In deep learning, computer vision is invading into natural language processing. With language processing, there is a continuous push toward lower level inputs, with word level inputs reaching the mainstream, and research investigating moving toward character level inputs emerging. This is similar to how deep learning has brought computer vision all the way down to raw pixel level for representational learning. Moreover, many recently popularized concepts and methods such as attention, dilated convolution, and encoder - decoder architectures that are shared between computer vision and natural language processing have legacies in computer vision. Though we have yet to know how much more computer vision can invade into natural language processing, computer vision has already made a significant impact on natural language processing through deep learning.

1 Introduction

It had once been the case where there was not much resemblance between computer vision tasks and natural language processing tasks. After all, we think of a text and an image quite different from one another.

Many works on language processing focused on understanding semantics and syntactic structures of languages with research focusing on parts of speech tagging (Leech, Garside, & Atwell, 1983) (Church, 1988) (Toutanova, Klein, Manning, & Singer, 2003), language parsing (Marcus, Marcinkiewicz, & Santorini, 1993) (De Marneffe, MacCartney, Manning, et al., 2006), semantic role labeling (Gildea & Jurafsky, 2002) (Carreras & Màrquez, 2005) (Palmer, Gildea, & Kingsbury, 2005), coreferene resolution (McCallum & Wellner, 2005) (H. Lee et al., 2011), conditional random fields (Lafferty, McCallum, & Pereira, 2001), named entity recognition (Collins & Singer, 1999) (Tjong Kim Sang & De Meulder, 2003), and etc. While some of them used statistical approaches to code linguistic rules, many of the approaches taken to solve these problems still re-

quired domain knowledge to come up with the features to solve these problems. Also, more domain knowledge was required to design a system that appropriately made use of these extracted information to perform the necessary language understanding tasks such as text classification, text summarization, information retrieval, question answering, and etc.

The story was similar for computer vision as feature detection and feature engineering process for computer vision seemed to be no use for natural language processing tasks. Edge detection (Canny, 1986) (Perona & Malik, 1990), object detection (Yuille, Hallinan, & Cohen, 1992) (Lowe, 1999) (Viola & Jones, 2001), and other feature detection processes along with other image processing such as image segmentation (Shi & Malik, 2000) required computer vision domain knowledge where many of them involved applying mathematical transformations using various functions that had not much use for natural language processing.

However, in deep learning, this no longer seems to be true. Vision approaches are becoming pervasive in natural language processing. Once departing from explicit linguistic rule based approach, language processing came down to word level inputs (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) (Bahdanau, Cho, & Bengio, 2014) (Sutskever, Vinyals, & Le, 2014) (Kim, 2014), where deep learning models would implicitly pick up the necessary semantic and syntatical structures during representational learning. Now, there is a new effort that is trying to push the field to character level inputs (Zhang & LeCun, 2015) (Zhang, Zhao, & LeCun, 2015) (Kim, Jernite, Sontag, & Rush, 2016) (Chung, Cho, & Bengio, 2016) (Bojanowski, Grave, Joulin, & Mikolov, 2016) (J. Lee, Cho, & Hofmann, 2016), following the foot prints of computer vision research in deep learning, which has pushed its input levels to all the way down to raw pixel level.

Moreover, many computer vision concepts and methods such as attention, dilated convolution and encoder - decoder architecture are making their way into natural language processing tasks. Even convolutional neural networks (CNNs) are moving into text modeling tasks.

CNNs have their foundations in vision, which started from a biological research (Hubel & Wiesel, 1968), followed by a neural network approach (Fukushima & Miyake, 1982), and a neural network approach with back propagation (LeCun et al., 1990) (LeCun, Bottou, Bengio, & Haffner, 1998) and had some early works on speech (Waibel, Hanazawa, Hinton, Shikano, & Lang, 1989) (LeCun, Bengio, et al., 1995), but they are quickly making a new mark on text and returning to speech again using the lessons learned from vision tasks. During a relatively short period of time, many promising and state of the art results have been produced in language processing using CNNs, many of which are direct applications of many lessons learned while performing vision tasks.

Overall, computer vision has made a significant impact on natural language processing through deep learning and time will tell how much more contribution computer vision can make on natural language processing.

2 Convolutional Neural Networks

Convolutions and sub-samplings make CNNs distinct from artificial neural networks (ANNs) (LeCun et al., 1995). Convolution is a process that rely on receptive filters to extract out local features. As an example, if an image is given as an input, a receptive filter can be thought of as a small sized filter with weights that slides through an image and computes a dot product with the filter weights and the portion of image pixels that lie beneath the filter. This computed value becomes a feature. While sliding through an image, a receptive filter will compute multiple values, and thus will create a feature map that consists of these multiple values or features. The weights of a receptive filter are parameters and can be learned using backpropagation. Thus each receptive filter becomes analogous to a template that learns different feature map and more receptive filters would create more feature maps that may potentially extract out more useful representations and information.

Sub-sampling is created to reduce the dimensions of a feature map. It helps to keep only the most important features and discard less important features within a feature map. Sub-sampling also allows generalization as emphasis is placed on the most notable features. Thus, sub-sampling allows to reduce computations complexity, while it helps to generalize the overall representations. Max pooling is a popular sub-sampling method, which extracts the max value within a fixed size filter that scans a feature map.

Stacking layers of convolutions and sub-sampling ultimately creates a hierarchical representation that relies on local features. From the bottom, local features are extracted, and these local features are combined to create higher level features that ultimately help to understand the actual object (LeCun et al., 1995).

3 Historical Context of Vision Invasion

Computer vision has a relatively longer history in deep learning than natural language processing does. Many of the early real world applications of neural networks were in vision tasks (Fukushima & Miyake, 1982) (LeCun et al., 1990) (LeCun et al., 1998) (Hinton & Salakhutdinov, 2006). While there were also language modeling work starting in early 2000s using RNNs (Lawrence, Giles, & Fong, 2000) (Bengio, Ducharme, Vincent, & Jauvin, 2003), they were still an emerging field.

It is also important to note that a major event that has pushed deep learning to today's popularity can be contributed to the 2012 ImageNet competition (Krizhevsky, Sutskever, & Hinton, 2012), which can be supported from the high number of citations that the work was able to garner in such a short period of time. Thus, deep learning has many legacies rooted in the vision.

Many of the efforts, as a result, initially focused on vision. Thus it is a natural consequence that deep learning researchers started to see if vision techniques could be applied to natural language processing, which ignited the vision invasion to language processing.

4 Move Towards Lower Level Input for Text Processing

In deep learning, the first significant move toward a lower level input was toward word level inputs with the help of word embeddings. Word embeddings got popularized with the introduction of Word2Vec (Mikolov et al., 2013). Word embeddings encode words to distributed numerical real value vectors where similar words are translated into vectors that are close to one another. Word embeddings can be trained using the context words around the target word from the provided corpus. The effectiveness of word embeddings was proven by many research results. Word level inputs alone without any semantic or syntactical structures as explicit inputs were enough to perform many different natural language processing tasks.

In text classification, some simple CNN architecture with word level inputs alone without any explicitly encoded semantic or syntactical structures seemed to work surprisingly well (Johnson & Zhang, 2014) (Kim, 2014). While the overall architecture was relatively simple as it involved one convolution layer on sentences and one max pooling layer with words encoded using word embeddings, a CNN based model achieved state of the art results in a few data sets for sentimental analysis and question classification (Kim, 2014).

For machine translation, RNN based architectures were able to outperform phrase-based statistical machine translation, using the word level inputs alone that are built on top of word embeddings (Sutskever et al., 2014).

Nowadays, word embeddings have become the mainstream in natural language processing tasks using deep learning. Many models for text classification (Kim, 2014) , machine translation (Bahdanau et al., 2014) (Sutskever et al., 2014) (Wu et al., 2016), text summarization (Rush, Chopra, & Weston, 2015), and image captioning (Karpathy & Fei-Fei, 2015) all utilize word embeddings.

Yet there are still some limitations with word level inputs that are trained on word embeddings. Especially out of vocabulary words make models that use word embeddings to perform poorly (Sutskever et al., 2014). Thus, modeling morphologically rich languages becomes a challenge.

Some notable research on character level based text classification were performed using CNN models (Zhang & LeCun, 2015) (Zhang et al., 2015). These CNN models relied solely on the character level text inputs. These works provided strong empirical evidence texts can be understood without much of semantics

or syntactic structures (Zhang & LeCun, 2015) and that language can be also understood from a very low character level (Zhang et al., 2015), just like how images are understood in terms of pixels.

Character level language modeling also produced promising results. Using perplexity as the evaluation metric, a character level based model performed on par with the state of the art results on the English Penn Tree Bank with far less parameters, while outperforming on morphologically rich languages against word-level LSTM baseline and analysis revealed that character level based models can internally learn semantics and orthographic information (Kim et al., 2016).

At a similar time, a character level input based embeddings was introduced (Bojanowski et al., 2016), where on of the core contributing researchers had been also behind the work of Word2Vec (Mikolov et al., 2013) . The concept of character level embeddings was to build word embeddings by summing character n-grams vectors of the target word. This helps to create better embeddings for rare words and also helps to deal with out of vocabulary words. It outperformed Word2Vec in many of the word similarity benchmark data sets.

Character level inputs also gave promising results in machine translation. Some of the researchers who made a significant contribution to popularizing Neural Machine Translation (NMT) came out with a character level decoder based NMT (Chung et al., 2016) and fully character level encoder - decoder based NMT (J. Lee et al., 2016) that incorporate and argued for character level machine translation system over word level machine translation system. Apart from resolving out of vocabulary words, character level inputs also remove text segmentation, which is a non trivial task (Chung et al., 2016) that can also potentially inject unnecessary bias and knowledge (J. Lee et al., 2016). As a result, with character level input, text processing can become truly end-to-end just like how image tasks are performed end-to-end and give superior results.

Yet, character level inputs also have drawbacks and it is an active research area to resolve these drawbacks. First, the training time becomes expensive (Luong & Manning, 2016) as now the input level is in characters rather than in words, which makes sequences extremely long. Second, it is more difficult for character level inputs based models to capture long term dependencies (Bojanowski, Joulin, & Mikolov, 2015). Also, some performance issues also exist where there are cases where word embeddings still perform better (?, ?). Yet the overall trend seems to indicate that the performance gap is closing down and we are starting to see cases where character level models outperform word level models, as more research effort is put into character level modeling (Kim et al., 2016) (Bojanowski et al., 2016) (Chung et al., 2016) (J. Lee et al., 2016).

While many early deep learning researchers were able to identify that both speech tasks and image tasks can boil down to pattern recognition (LeCun et al., 1995), big discrepancies between approaches toward texts and images had put a barrier between computer vision and natural language processing.

However with vision invading into text, this gap closing down.

5 Computer Vision Concepts and Methods into Natural Language Processing

This section examines a few concepts and methods that have legacies in computer vision, but got recently popularized in natural language processing.

5.1 Attention

The inspiration of attention came from how humans process an image. Humans tend to scan an image to locate the important region and start focusing on that region rather than paying attention to every single region and detail. Some early works on neural networks model for attention can be traced back for a image classification task (Larochelle & Hinton, 2010), where the concept of attention was applied by creating a neural network that would sequentially accumulate features from a few fixated positions of an image. The concept of attention was also applied to generating synthesized handwriting, given a sequence of characters as inputs (Graves, 2013). The core idea was to use a weighted sum of the input characters for generating each output vector, which would create a distribution that will emphasize a few important characters that are responsible for generating each output vector. This work could be viewed as the first adoption of attention into text. Many image models also adopted attention mechanism (Mnih, Heess, Graves, et al., 2014) (Ba, Mnih, & Kavukcuoglu, 2014) for image classification tasks.

However, attention mechanism was popularized in natural language processing when (Bahdanau et al., 2014) effectively integrated attention that is a variation of (Graves, 2013) into machine translation, which helped the system to attend to the necessary words of a sentence when performing machine translation. Soon speech recognition also adopted attention method (Chorowski, Bahdanau, Serdyuk, Cho, & Bengio, 2015). Now attention method is popular in both computer vision (Ba et al., 2014) (Sermanet, Frome, & Real, 2014) (Chen et al., 2015) (Xu et al., 2015) and natural language processing (Bahdanau et al., 2014) (Luong, Pham, & Manning, 2015) (Rush et al., 2015) (Chan, Jaitly, Le, & Vinyals, 2015) (Bahdanau, Chorowski, Serdyuk, Brakel, & Bengio, 2016)

5.2 Dilated Convolution

Dilated convolution was introduced to capture multi-scale context in image segmentation (Yu & Koltun, 2015). The idea was to pad zeros in between columns and rows of receptive filters, which would increase the receptive filter size and

better capture the global context, without increasing the number of parameters. From an image perspective, as receptive filter size becomes bigger, the filter can capture pixels that are farther apart, which helps to capture more global context. At the same time, as zeros, which are not parameters, were padded in between the filter columns and rows to increase the filter size, the computation does not become expensive for backpropagation during training. Interestingly, this turned out to be useful for natural language processing too as the inability to capture global context was one of the bottlenecks for using CNNs for natural language processing .

A notable work outside of vision to adopt dilated convolution was audio generation. WaveNet (Oord et al., 2016) was able to generate audio sound that closely replicate natural human voice, and outperformed all previous existing speech synthesis systems.

Adoption of dilated convolution into machine translation soon followed. ByteNet (Kalchbrenner et al., 2016) was able to create a linear time running machine translation using dilated convolution that gave state of the art result in character to character machine translation.

Recently, dilated convolution was also introduced in entity recognition (Strubell, Verga, Belanger, & McCallum, 2017) and language parsing (Strubell & McCallum, 2017).

5.3 Encoder and Decoder Architectures

The terms encoding and decoding in deep learning can be understood from a statistics context. Encoding can be understood as creating a latent variable based on observed variables, which is an inference step, while decoding is reversing the above described process, which can be viewed as a generative step.

The first work that successfully incorporates the concepts of encoding and decoding into a deep neural network can be traced back to a decade early work (Hinton & Salakhutdinov, 2006), which dealt with dimensionality reduction and reconstruction on images (in other words image compression and decompression). The idea was to stack layers of artificial neurons with each layer decreasing artificial neuron numbers. This will ultimately "encode" and create a latent representation with a reduced dimension. Once having created a latent representation, another stacks of layers of artificial neurons with each layer increasing artificial neuron number will be constructed, which will "decode" the latent representation to the desired target representation, which in this case actually becomes the input representation or the original image (Strictly speaking, this encoder - decoder architecture is often referred to as autoencoder, which can be understood as a more specific case of a general encoder - decoder architecture).

Initially, these types of encoder - decoder architectures were widely used for

understanding the representational learning process of neural networks along with applications in dimensionality reduction and reconstruction. They also influenced a few later introduced models such as Varitional AutoEncoder (VAE) (Kingma & Welling, 2013) or Generative Adversial Network (GAN) (Goodfellow et al., 2014) that are widely used for image generations these days.

Yet this encoder - decoder architecture recently played a significant role in creating a machine translation system known as Neural Machine Translation (NMT). NMTs are based on encoder - decoder architectures that encode sequences into a latent vector and decode it to create translated sequences. (Kalchbrenner & Blunsom, 2013) (Cho et al., 2014) (Bahdanau et al., 2014) (Sutskever et al., 2014)

What is also interesting to note is that while NMTs are now widely associated with RNN based encoder - decoder architecture (Cho et al., 2014) (Bahdanau et al., 2014) (Sutskever et al., 2014), the very early work actually utilized a CNN based model for the encoder process that created the latent vector (Kalchbrenner & Blunsom, 2013).

In the end, the legacies of many encoder - decoder architectures that have become popular in natural language processing have roots in computer vision research.

5.4 Residual Connections

ResNet is a popular deep learning model for image classification due to extremely low error rate (He, Zhang, Ren, & Sun, 2016). They were able to achieve such a low error rate by creating an extremely deep CNN model without suffering from the vanishing gradient problem. They approached the vanishing gradient problem by creating shortcut connections (residual connections) that connect layers that are not adjacent, which allowed gradients to flow more easily to deeper layers. This allowed the architecture to rely on an extremely deep stacked layers. Recently, a machine translation system that is solely based on CNNs (Gehring, Auli, Grangier, Yarats, & Dauphin, 2017) was introduced, which adopted residual connections. The machine translation system was able to achieve state of the art results on multiple benchmark data sets. Overall many CNNs based models have learned that deep networks perform extremely well in practice through computer vision tasks (Krizhevsky et al., 2012) (Simonyan & Zisserman, 2014) (Szegedy et al., 2015) (He et al., 2016), and this important lesson learned from computer vision is being transferred to natural language processing (Conneau, Schwenk, Barrault, & Lecun, 2016) (Gehring et al., 2017)

6 Future Research Directions

Many of the directions as to where natural language processing research is headed to in deep learning can be understood through examining how computer vision research has progressed and where it is headed to. One notable area right now would be research that focuses on understanding and developing character level model for language understanding. As natural language processing is not yet truly end-to-end, which is in contrast to how most of computer vision research is done in deep learning, many deep learning researchers seem to be pushing for truly end-to-end approach for natural language processing, where character level modeling seems like one appropriate approach.

In terms of one particularly active research area in computer vision that has not been yet effectively integrated into natural language processing would be generative models. Image generation using generative models such as generative adversarial networks (GANs) or variational autoencoders (VAEs) is still an active research area in computer vision. Adopting similar methods in novel ways to text and speech generations may potentially lead to fruitful contributions to natural language processing.

At the same time, it is important to not forget that many recently popularized methods are built on top of the works that had been introduced earlier, where some of them are decades old. While the deep learning community tends to push for new and trendy topics, many prominent researchers seem to be still getting inspirations from old ideas out there that can potentially be adopted in a new and creative way and result in a significant advancement in deep learning research. In a similar manner, it is also interesting to realize that a novel integration of an already existing work to a new field such as applying attention and encoder - decoder architecture to texts also provides significant advancement in research.

7 Conclusion

This survey paper showed how computer vision is invading into natural language processing. From a historical context, deep learning research has a relatively longer history in vision and vision tasks played a critical role in deep learning to gain today's popularity, which gives an intuition as to why this might be the case. With language processing, there is a continuous push toward lower level inputs, with word level inputs reaching the mainstream, and research investigating moving toward character level inputs emerging. This is similar to how deep learning has brought computer vision all the way down to raw pixel level for representational learning. Many concepts and methods such as attention, dilated convolutions, encoder - decoder architectures, and residual connections that are adopted by natural language processing tasks all have legacies in computer vision. These outcomes show that in deep learning, the progress in computer

vision research may hold a key as to how natural language processing research would move forward from where it is now.

References

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *Acoustics, speech and signal processing (icassp), 2016 ieee international conference on* (pp. 4945–4949).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137–1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bojanowski, P., Joulin, A., & Mikolov, T. (2015). Alternative structures for character-level rnns. *arXiv preprint arXiv:1511.06303*.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*(6), 679–698.

Carreras, X., & Màrquez, L. (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning* (pp. 152–164).

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., & Nevatia, R. (2015). Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. In *Advances in neural information processing systems* (pp. 577–585).

Chung, J., Cho, K., & Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on applied natural language processing* (pp. 136–143).

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In *1999 joint sigdat conference on empirical methods in natural language processing and very large corpora.*

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of lrec* (Vol. 6, pp. 449–454).

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets* (pp. 267–285). Springer.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, *28*(3), 245–288.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215–243.

Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In *Emnlp* (Vol. 3, p. 413).

Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. v. d., Graves, A., & Kavukcuoglu, K. (2016). Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3128–3137).

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *Aaai* (pp. 2741–2749).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Larochelle, H., & Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order boltzmann machine. In *Advances in neural information processing systems* (pp. 1243–1251).

Lawrence, S., Giles, C. L., & Fong, S. (2000). Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, *12*(1), 126–140.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* (pp. 396–404).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task* (pp. 28–34).

Lee, J., Cho, K., & Hofmann, T. (2016). Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*.

Leech, G., Garside, R., & Atwell, E. S. (1983). The automatic grammatical tagging of the lob corpus. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, *7*, 13–33.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. the proceedings of the seventh ieee international conference on* (Vol. 2, pp. 1150–1157).

Luong, M.-T., & Manning, C. D. (2016). Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv preprint arXiv:1604.00788*.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, *19*(2), 313–330.

McCallum, A., & Wellner, B. (2005). Conditional models of identity uncertainty with application to noun coreference. In *Advances in neural information processing systems* (pp. 905–912).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212).

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*(1), 71–106.

Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence*, *12*(7), 629–639.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Sermanet, P., Frome, A., & Real, E. (2014). Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(8), 888–905.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Strubell, E., & McCallum, A. (2017). Dependency parsing with dilated iterated graph cnns. *arXiv preprint arXiv:1705.00403*.

Strubell, E., Verga, P., Belanger, D., & McCallum, A. (2017). Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2660–2670).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on natural language learning at hlt-naacl 2003-volume 4* (pp. 142–147).

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1* (pp. 173–180).

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer vision and pattern recognition, 2001. cvpr 2001. proceedings of the 2001 ieee computer society conference on* (Vol. 1, pp. I–I).

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, *37*(3), 328–339.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057).

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yuille, A. L., Hallinan, P. W., & Cohen, D. S. (1992). Feature extraction from faces using deformable templates. *International journal of computer vision*, *8*(2), 99–111.

Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657).