WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Study of Deep Learning and NLP

Abhishek A. Gulhane

Research Scholar,JJT University, Jhunjhunu, Rajastan, India

ABSTRACT

Alert is an increasingly popular medium used in a wide range of neural architectures. The medium itself has been realized in a variety of formats. Still, because of the fast-paced advances in this sphere, a regular overview of attention is still missing. In this composition, we define a unified model for attention architectures in natural language processing, with a focus on those designed to work with vector representations of the textual data. We propose a taxonomy of attention models according to four confines the representation of the input, the harmony function, the distribution function, and the assortment of the input and/ or affair. We present the samples of how former information can be exploited in attention models and bandy ongoing disquisition sweats and open challenges in the area, furnishing the first extensive categorization of the vast body of literature in this provocative sphere.

Arguably, effective results to analogous problems should factor in a notion of connection, so as to concentrate the computational resources on a confined set of important rudiments. One possible approach would be to conform results to the specific Order at hand, in order to more exploit known disagreement of the input, by point engineering. For illustration, in the argumentative analysis of conclusive essays, one could decide to give special emphasis to the final judgment. Still, analogous an approach is not always doable, especially if the input is long or truly information-rich, analogous as in text summarization, where the affair is the condensed interpretation of a possibly lengthy

Over the last several times, the field of natural language processing has been propelled forward by an explosion in the use of deep knowledge models. This composition provides a brief prolusion to the field and a quick overview of deep literacy Architectures and styles. It also sifts through the plethora of recent studies and summarizes a large assortment of applicable contributions. Analyzed disquisition areas include several core verbal processing issues in addition to multitudinous operations of computational linguistics. A discussion of the current state of the art is also handed along with recommendations for future disquisition in the field.

Keywords:Computational linguistics, natural language processing (NLP), Deep Learning, neural networks(NN)

1. Introduction

In Measure problems that involve the processing of natural language, the rudiments composing the source text are characterized by having each a different connection to the task at hand. For case, in aspect- predicated sentiment analysis, cue words, analogous as “ good” or “ bad,” could be applicable to some aspects under consideration, but not to others. In machine paraphrase, some words in the source text could be irrelevant in the paraphrase of the coming word. In a visual question- answering task, background pixels could be irrelevant in answering a question regarding an object in the focus but applicable to questions regarding the scenery.

Arguably, effective results to analogous problems should factor in a notion of connection, so as to concentrate the computational resources on a confined set of important rudiments. One possible approach would be to conform results to the specific Order at hand, in order to more exploit known disagreement of the input, by point engineering. For illustration, in the argumentative analysis of conclusive essays, one could decide to give special emphasis to the final judgment. Still, analogous an approach is not always doable, especially if the input is long or truly information-rich, analogous as in text summarization, where the affair is the condensed interpretation of a possibly lengthytext sequence.

* *Corresponding author.* Tel.: +919890398070;
E-mail address: aagabhi@gmail.com

Another approach of adding popularity amounts to machine learning the connection of input rudiments. In that way, neural architectures could automatically weigh the connection of any region of the input and take such a weight into account while performing the main task. The commonest result to this problem is a medium known as Alert.

NLP enables computers to perform various language related tasks, ranging from sentiment analysis to parsing and machine paraphrase. In this composition, we review the history of the NLP disquisition trend, which shifted from rule- predicated methodology supported by Chomsky's verbal propositions to the statistics- predicated approach that are reckoned on statistics, information proposition, and machine knowledge. The recent trends in deep knowledge- predicated NLP are also mooted. Over the formerly numerous decades, the maturity of NLP problems that reckoned on shallow built models, analogous as SVM and logistic regression, were trained on stingy and greatly high dimensional features. Following the trend of the increasingly advanced deep knowledge models and algorithms, the recent NLP inquiries have concentrated more on the operation of new deep knowledge architectures. Generally, numerous complex neural networks predicated on thick point representations have performed results superior to traditional Styles in working certain NLP tasks. This trend is sparked by the successful operation of word embedding, a distributional vector that can be used to capture and measure the similarity between word vectors. This work also introduces other deep Knowledge related models applied to NLP tasks, analogous as convolutional neural networks (CNNs), intermittent neural networks (RNNs), and attention medium.

THE field of natural language processing (NLP) encompasses a variety of motifs, which involves the computational processing and understanding of mortal languages. Since the 1980s, the field has increasingly reckoned on data- driven Computation involving statistics, probability, and machine knowledge (1), (2). Recent increases in computational power and parallelization, exercised by graphical processing units (GPUs) (3), (4), now allow for “deep knowledge,” which utilizes artificial neural networks (ANNs), sometimes with billions of trainable parameters (5). In addition, the contemporary vacuity of large data sets, eased by sophisticated data collection processes, enables the training of analogous deep Architectures (6) – (8).

The motifs of NLP and AI, including deep knowledge, are introduced in Part II. The ways in which deep knowledge has been used to break problems in core areas of NLP are presented in Part III. The Part is broken down into several sub Parts, videlicet natural language modeling (Part III-A), morphology (Part III-B), parsing (Part III-C), and semantics (Part III-D). Operations of deep knowledge to further practical areas are mooted in Part IV. Specifically mooted are information recovery (IR) (Part IV-A), information birth Part (IV-B), text type (Part IV-C), text generation (Part IV-D), summarization (Part IV-E), question answering (QA) (Part IV-F), and machine paraphrase (Part IV-G). Conclusions are also drawn in Part V with a brief summary of the state of the art as well as prognostications, suggestions, and other studies on the future of this roundly evolving area.

2. Recap of NLP &Deep Learning

In this part, significant issues that draw the attention of researchers and practitioners are introduced, followed by a brisk explanation of the deep knowledge architectures generally used in the field.

A. Natural Language Processing:-

The field of NLP, also known as computational linguistics, involves the engineering of computational models and processes to break practical problems in understanding mortal languages. These results are used to make useful software. Work in NLP can be divided into two broad subareas core areas and operations, although it's sometimes delicate to distinguish fluently to which areas issues belong. The core areas address fundamental problems analogous as language modeling, which underscores quantifying associations among naturally being words; morphological processing, dealing with segmentation of meaningful factors of words and relating the true corridor of speech (POSs) of words as used; syntactic processing, or parsing, which builds judgment plates as possible precursors to semantic processing; andsemantic processing, which attempts to distill meaning of words, expressions, and advanced position factors in text. The operation areas involve motifs, analogous as birth of useful information (e.g., named realities and relations), paraphrase of Textbook between and among languages, summarization of written factory, automatic answering of questions by inferring answers, and type and clustering of documents. Constantly, one needs to handle one or further of the core issues successfully and apply those ideas and procedures to break practical problems. Presently, NLP is primarily a data- driven field using statistical and probabilistic computations along with machine knowledge. In the history, machine knowledge approaches, analogous as naïve Bayes, k-nearest neighbors, hidden Markov models, conditional arbitrary fields (CRFs), decision trees, arbitrary timbers, and support vector machines, were considerably used. Still, during the formerly several times, there has been a commercial transformation, and these approaches have been entirely replaced, or at least enhanced, by neural models, mooted next.

B. NN &Deep Learning

Neural networks are composed of connected bumps, or neurons, each entering some number of inputs and supplying an affair. Each of the bumps in the affair layers performs weighted sum computation on the values they admit from the input Bumps and also induce labors using simple nonlinear transformation functions on these summations. Corrections to the weights are made in response to individual crimes or losses that the networks cortege at the affair bumps. Analogous corrections are generally made in modern networks using stochastic grade descent, considering the derivatives of crimes at the bumps, an approach called backpropagation (13). The main factors that distinguish different types of networks from each other are how the bumps are connected and the number of layers. Basic networks in which all bumps can be organized into successive layers, with every knot entering inputs only from bumps in earlier layers, are known as feedforward neural networks (FFNNs). While there is no clear agreement on exactly what defines a DNN, generally, networks with multiple sheltered layers are considered deep and those with multitudinous layers are considered truly deep (7).

i. Convolutional Neural Networks Convolutional neural networks (CNNs) (14), (15), erected upon Fukashima’sneocognitron (16), (17), decide the name from the complication operation in mathematics and signal processing. CNNs use functions, known as adulterants, allowing for simultaneous analysis of different features in the data (18), (19). CNNs are vastly used in image and video processing, as well as speech and NLP (20) – (23). Constantly, it is not important precisely where certain features do, but rather whether or not they appear in particular points. Therefore, pooling operations can be used to minimize the size of point maps (the labors of theconvolutional adulterants). The sizes of analogous pools are generally small to help the loss of too important perfection.

ii. Recursive Neural Networks Important like CNNs, recursive networks (24), (25) use a form of weight sharing to minimize training. Still, whereas CNNs partake weights horizontally (within a caste), recursive nets partake weights vertically (between layers). This is particularly fascinating, as it allows for easy modeling of structures analogous as parse trees. In recursive networks, a single tensor (or a generalized matrix) of weights can be used at a low position in the tree and also used recursively at successively advanced situations (26).

iii. Recursive Neural Networks and Long Short- Term Memory Networks A type of recursive neural network that has been used heavily is the intermittent neural network (RNN) (27), (28). Since important of NLP is dependent on the order of words or other rudiments analogous as phonemes or rulings, it's useful to have memory of the former rudiments when recovering new bones (29) – (31). Sometimes, backward dependences live, i.e., correct processing of some words may depend on words that follow. Thus, it's salutary to look at rulings in both directions, forward and backward, using two RNN layers and combining their labors. This arrangement of RNNs is called a bidirectional RNN. It may also lead to a better final representation if there is a sequence of RNN layers. This may allow the effect of an input to crawl longer than a single RNN caste, allowing for longer term goods. This setup of successive RNN cells is called an RNN mound (32), (33).

iv. Attention Mechanisms and Transformer For tasks analogous as machine paraphrase, text summarization, or captioning, the affair is in textual form. Generally, this is done through the use of encoder – decoder couples. A garbling ANN is used to produce a vector of a particular length and a decoding ANN is used to return variable- length text predicated on this vector. The problem with this scheme, which is shown inFig. 1, is that the RNN is forced to render an entire sequence to a finite length vector, without regard to whether or not any of the inputs are more important than others.

A robust result to this is that of attention. The first noted use of an attention medium (38) used a thick caste for annotated weighting of an RNN’s sheltered state, allowing the network to learn what to pay attention to in Agreement with the current sheltered state and reflection. Such a medium is present inFig. 1 . Variants of the medium have been introduced, popular bones including convolutional (39), intertemporal (40), restarted (41), and self-attention (42). Tone- attention involves furnishing attention to words in the same judgment. For illustration, during garbling a word in an input judgment, it's salutary to project variable amounts of attention to other words in the judgment. During Decoding to produce a performing judgment, it makes sense to give applicable attention to words that have formerly been produced. Tone- attention, in particular, has come considerably used in a state-of-the- art encoder – decoder model called motor (42). The motor model, shown inFig. 2, has a number of encoders and decoders piled on top of each other, tone-attention in each of the encoder and decoder units, and cross attention between the encoders and the decoders. It uses multiple cases of attention in resembling and eschews the use of recurrences and complications. The motor has come a definitive element in utmost state-of-the- art neural networks for NLP.

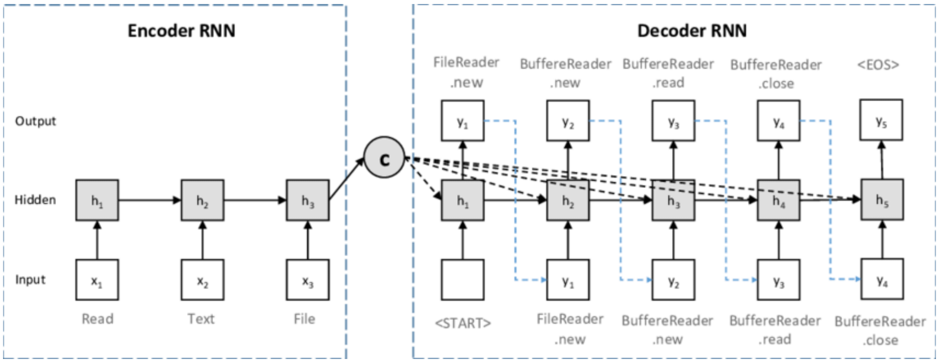

Fig 1:- Encoder–decoder architectures

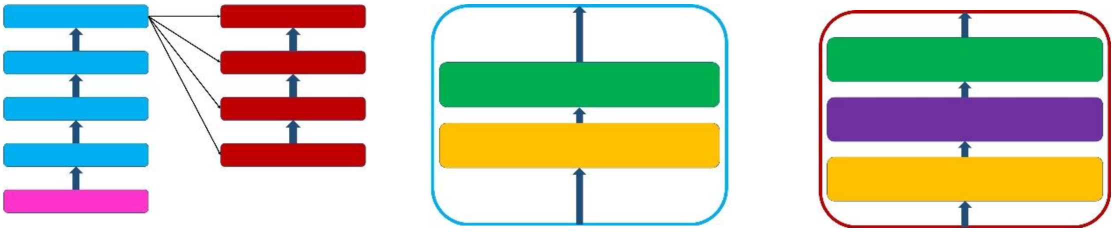

Fig. 2. Transformer model. (a) Transformer with four “encoders” followed by four “decoders,” all following a “positional encoder.” (b) Inner workings of each “encoder,” which contains a self-attention layer followed by a feed forward layer. (c) Inner workings of each “decoder,” which contains a self-attention layer followed by an attentional encoder–decoder layer and then a feed forward layer.

v. Residual Connections and Dropout In deep networks, trained via backpropagation (13), the slants used to correct for error constantly dematerialize or explode (43). This can be eased by choosing activation functions, analogous as the remedied direct unit (ReLU) (44), which do not cortege regions that are acratically steep or have basically small slants. Also, in response to this issue, as well as others (45), residual connections are constantly used. Analogous connections are simply those that skip layers (generally one). Still, this cuts in if used in everyinterspersinglayer.half the number of layers through which the grade must backpropagate. Such a network is known as a residual network (ResNet). A number of variants live, including trace networks (46) and DenseNets (47). Another important system used in training ANNs is hustler. In hustler, some connections and maybe indeed bumps are killed, generally erratically, for each training batch (small set of samples), varying which bumps are killed each batch. This forces the network to distribute its memory across multiple paths, helping with generality and lessening the liability of overfitting to the training data.

3. Deep Learning In Core Areas Of NLP

The core issues are those that are constitutionally present in any computational verbal system. To perform paraphrase, text summarization, image captioning, or any other verbal task, there must be some understanding of the bolstering language. This understanding can be broken down into at least four main areas language modeling, morphology, parsing, and semantics.

Language modeling can be viewed in two ways. First, it determines which words follow which. By extension, still, this can be viewed as determining what words mean, as individual words are only weakly meaningful, inferring their full value only from their relations with other words. Morphology is the study of how words themselves are formed. It considers the roots of words and the use of prefixes and suffixes, mixes, and other intraword bias, to display tense, gender, plurality, and another verbal constructs. Parsing considers which words modify others, forming constituents, leading to a sentential structure. The area of semantics is the study of what words mean. It considers the meanings of the individual words and how they relate to and modify others, as well as the surrounds these words appear in and some degree of world knowledge, i.e., “common sense.” There is a significant amount of overlap between each of these areas. Therefore, multitudinous models analyzed can be classified as belonging in multiple sections. As analogous, they are mooted in the most applicable sections with logical connections.

i. Language Modeling and Word Embeddings

Arguably, the most important task in NLP is that of language modeling. Language modeling is an essential piece of nearly any operation of NLP. Language modeling is the process of creating a model to predict words or simple verbal factors given former words or factors (48). This is useful for operations in which a user types input, to give predictive capability for fast text entry. Still, its power and versatility radiate from the fact that it can implicitly internee syntactic and semantic connections among words or factors in a direct neighborhood, making it useful for tasks analogous as machine paraphrase or text summarization. Using Vaticination, analogous programs can induce further applicable, mortal-sounding rulings.

1) Neural Language Modeling A problem with statistical language models was the incapacity to deal well with Synonyms or out-of-vocabulary (OOV) words that weren't present in the training corpus. Progress was made in working the problems with the prolusion of the neural language model (49). While important of NLP took another decade to begin to use ANNs heavily, the language modeling communitydirectly took advantage of them and continued to develop sophisticated models, multitudinous of which were epitomized by De Mulder etal. (50).

2) Evaluation of Language Models While neural networks have made Advancements in the language modeling field, it's hard to quantify advancements. It's desirable to estimate language models independently of the operations in which they appear. A number of criteria have been proposed, but no perfect result has yet been factory. (51) – (53) the most generally used metric is confusion, which is the inverse probability of a test set formalized by the number of words. Confusion is a reasonable dimension for language modelings trained on the same data sets, but when they are trained on different vocabularies, the metric becomes lower meaningful. Luckily, there are several standard data sets that are used in the field, allowing for comparison. Two analogous data sets are the Penn Treebank (PTB) (54) and the Billion Word Benchmark (55).

3) Memory Networks and Attention Mechanisms in Language Modeling Daniluketal. (56) Tested several networks using variations of attention mechanisms. The first network had a simple attention medium, which was not fully connected, having a window length of five. They presupposed that using a single value to predict the coming honorary, render information for the attentional unit, and crack the information in the attentional unit hinders a network, as it's delicate to train a single parameter to perform three distinct tasks simultaneously. Therefore, in the alternate network, they designed each Knot to have two labors one to render and crack the information in the attentional unit, and another to predict the coming commemoratives explicitly. In the third network, they further separated the labors, using separate values to render the information entering the attentional unit and crack the information being reacquired from it. Tests on a Wikipedia corpus showed that the attention medium bettered confusion compared to the birth and that successively adding the alternate and third parameters led to further increases. It was also noted that only the former five or so commemoratives carried important value (hence the selection of the window size of five). Therefore, they tested a fourth network that simply used residual connections from each of the former five units. It was factory that this network also handed results analogous to multitudinous larger RNNs and LSTMs, suggesting that reasonable results can be achieved using simpler networks.

4) Summary of Core Issues Deep knowledge has generally performed truly well, surpassing being countries of the art in multitudinous individual core NLP tasks and has thus created the foundation on which useful natural language Operations can and are being erected. Still, it's clear from examining the disquisition reviewed also that natural language is an enigmatically complex content, with myriad

core or introductory tasks, of which deep knowledge has only grazed the face. It's also not clear how architectures for competently executing individual core tasks can be synthesized to make a common edifice, possibly a much more complex distributed neural architecture, to show capability in multiple or "all" core tasks. Farther constitutionally, it's also not clear, how literacy of introductory tasks, may lead to superior performance in applied tasks, which are the ultimate engineering pretensions, especially in the terrain of structure effective and effective deep Knowledge models. Multitudinous, if not most, successful deep knowledge architectures for applied tasks, mooted in Section IV, feel to hesitate unambiguous architectural factors for core tasks and learn analogous tasks implicitly. Thus, some researchers argue that the connection of the large amount of work on core issues is not fully justified, while others argue that further extensive disquisition in analogous areas is necessary to more understand and develop systems which more perfectly perform these tasks, whether explicitly or implicitly.

4. Operations OF NLP USING DEEP Learning

Knowledge While the study of core areas of NLP is important to understanding how neural models work, it's meaningless in and of itself from an engineering perspective, which values the Operations that profit humanity, not pure philosophical and scientific inquiry. Current approaches to working several directly useful NLP tasks are epitomized also. Note that the issues included also are only those involving the processing of text, not the processing of verbal speech. Because speech processing (162), (163) requires moxie on several other motifs including audial processing, it's generally considered another field of its own, sharing multitudinous parallels with the field of NLP. The number of studies in each mooted area over the last decade is shown in Fig. 3

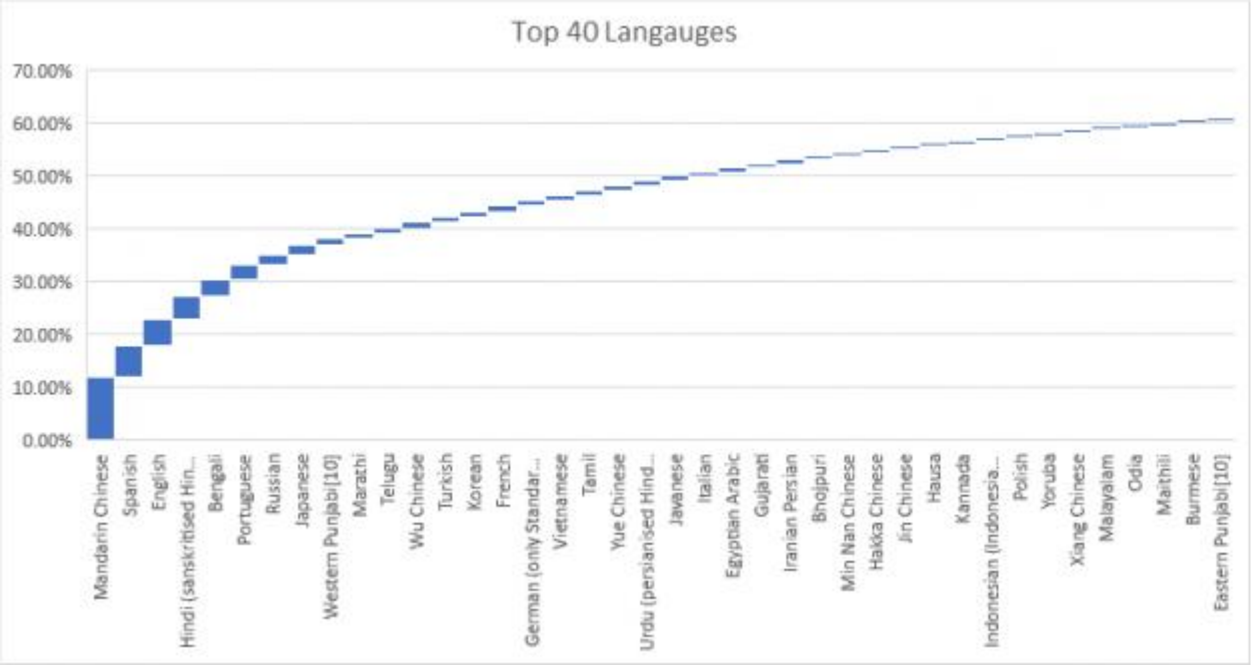

Fig 3: NLP bandied have witnessed growth in recent times, with the largest growth

4.1 Information Retrieval

The purpose of IR systems is to help people find the right (most useful) information in the right (most accessible) format at the right time (when they need it) (164). Among multitudinous issues in IR, a primary problem that needs addressing pertains to ranking documents with respect to a query string in terms of connection scores for ad hoc recovery tasks, similar to what happens in a quest machine. Deep Knowledge models for ad hoc recovery match handbooks of queries to handbooks of documents to gain connection scores. Thus, analogous models have to concentrate on producing representations of the relations among individual words in the query and the documents. Some representation-concentrated approaches make deep knowledge models to produce good representations for the handbooks and also match the representations directly (133), (165), (166), whereas interactionfocused approaches first Make original relations directly and also use DNNs to learn how the two pieces of text match predicated on word relations (133), (167), (168). When matching a long document to a short query, the applicable portion can potentially do anywhere in the long document and may also be distributed, thus, chancing how each word in the query relates to portions of the document is helpful. Apprehensive of the specific conditions for IR, Guoetal. (169) Erected a neural architecture called DRMM, enhancing an commerce- concentrated model that feeds quantized histograms of the original commerce intensities to an MLP for matching. In resemblant, the query terms go through a small subnetwork on their own to establish term significance and term dependences. The labors of the two similar networks are mixed at the top so that the connection of the document to the query can be better learned. DRMM achieved the state-of-the- art performance for its time.

5. Conclusion

Beforehand operations of NLP included a well- accredited but dewy-eyed algebra word problem solver program called Pupil, as well as interesting but severely constrained conversational systems analogous as Eliza, which acted as a “psychotherapist”, and another that conversed about manipulating blocks in a micro world. Presently, largely advanced operations of NLP are ubiquitous. These include Google’s and Microsoft’s machine translators, which translate more or less competently from a language to scores of other languages, as well as a number of bias which exercise voice commands and respond in like. The emergence of these sophisticated operations, particularly in posted settings, acts as a testament to the emotional accomplishments that have been made in this sphere over the last sixty or so times. Without a distrustfulness, implausible progress has taken place, particularly in the last several times. As has been shown, this recent progress has a clear unproductive relationship with the remarkable advances in ANNs. Considered an “old” technology just a decade ago, these machine learning constructs have steered in progress at an unknown rate, breaking performance records in myriad tasks in miscellaneous fields. In particular, deep neural architectures have inculcated models with advanced performance in natural language tasks, in terms of “amiss” criteria. Consolidating the analysis of all the models surveyed, a numerous general trends can be summarized. Both convolutional and intermittent samples had contributed to the state of the art in the recent history; still, of truly late, stacks of attention-powered motor units as encoders and constantly decoders have constantly produced superior results across the rich and varying terrain of the NLP field. These models are generally heavily pre-trained on general language knowledge in an unsupervised or supervised manner and kindly easily trained on specific tasks in a supervised fashion. Second, attention mechanisms alone, without recurrences or complications, feel to give the swish connections between encoders and decoders. Third, forcing networks to examine different features (by performing multiple tasks) generally improves results. Ultimately, while largely negotiating networks generally optimizes results, there is no cover for cultivating networks with large quantities of high- quality data, although pre-training on large general corpora seems to help immensely. Following from this final observation, it may be useful to direct farther disquisition trouble toward pre-training methodologies, rather than developing largely specialized Factors to squeeze the last drops of performance from complex models. While the numerous astral architectures being proposed each month are largely competitive, muddling the process of relating a winning architecture, the styles of evaluation used add just as important complexity to the problem. Data sets used to estimate new models are constantly generated specifically for those models and are also used only several farther times, Still, although consolidated data sets encompassing several fatal tasks analogous as Cement have started to crop. As the features and sizes of these data sets are largely variable, this makes comparison delicate. Utmost subfields of NLP, as well as the field as a whole, would benefit from extensive, large-scale exchanges regarding the necessary contents of analogous data sets, followed by the florilegium of analogous sets. In addition to high variability in evaluation data, there are numerous Criteria used to estimate performance on each task. Continually, comparing similar models is delicate because different criteria are reported for each. Agreement on particular sets of criteria would go a long way toward icing clear comparisons in the field.

References

[1] K. S. Jones, “Natural language processing: A historical review,” in Current Issues in Computational Linguistics: In Honour of Don Walker. Dordrecht, The Netherlands: Springer, 1994, pp. 3–16.

[2] E. D. Liddy, “Natural language processing,” in Encyclopedia of Library and Information Science, 2nd ed. New York, NY, USA: Marcel Decker, Inc., 2001.

[3] A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, and N. Andrew, “Deep learning with cots HPC systems,” in Proc. ICML, 2013, pp. 1337–1345.

[4] R. Raina, A. Madhavan, and A. Y. Ng, “Large-scale deep unsupervised learning using graphics processors,” in Proc. ICML, 2009, pp. 873–880.

[5] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep Learning, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[7] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural Netw., vol. 61, pp. 85–117, Jan. 2015.

[8] D. C. Ciresan et al., “Flexible, high performance convolutional neural networks for image classification,” in Proc. IJCAI, 2011, vol. 22, no. 1, p. 1237.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu,and P. Kuksa, “Natural language processing (almost) from scratch,”J. Mach. Learn.Res., vol. 12 pp. 2493–2537, Aug. 2011.

[10] Y. Goldberg, “Neural network methods for natural language processing,”Synth. Lect. Hum. Lang. Technol., vol. 10, no. 1, pp. 1–309, 2017.

[11] Y. Liu and M. Zhang, “Neural network methods for natural languageprocessing,” Comput. Linguistics, vol. 44, no. 1, pp. 193–195,Mar. 2018.

[12] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends indeep learning based natural language processing,” IEEE Comput. Intell.Mag.,vol. 13, no. 3, pp. 55–75, Aug. 2018.

[13] D. Rumelhart, G. Hinton, and R. Williams, “Learning internal representationsby error propagation,” UCSD, La Jolla, CA, USA, Tech. Rep.ICS-8506, 1985.

[14] Y. LeCun et al., “Backpropagation applied to handwritten zipcode recognition,” Neural Comput., vol. 1, no. 4, pp. 541–551,Dec. 1989.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-basedlearning applied to document recognition,” Proc. IEEE, vol. 86, no. 11,pp. 2278–2324, Nov. 1998.

[16] K. Fukushima, “Neocognitron: A self-organizing neural network modelfor a mechanism of pattern recognition unaffected by shift in position,”Biol. Cybern., vol. 36, no. 4, pp. 193–202, Apr. 1980.

[17] K. Fukushima and S. Miyake, “Neocognitron: A new algorithm forpattern recognition tolerant of deformations and shifts in position,”Pattern Recognit., vol. 15, no. 6, pp. 455–469, Jan. 1982.

[18] Y. LeCun et al., “Convolutional networks for images, speech, andtime series,” in The Handbook of Brain Theory and Neural Networks,vol. 3361, no. 10. Cambridge, MA, USA: MIT Press, 1995.

[19] A. Krizhevsky, “One weird trick for parallelizing convolutionalneural networks,” 2014, arXiv:1404.5997. [Online]. Available: http://arxiv.org/abs/1404.5997

[20] Y. Kim, “Convolutional neural networks for sentence classification,”2014, arXiv:1408.5882. [Online]. Available: http://arxiv.org/abs/1408.5882

[21] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutionalneural network for modelling sentences,” 2014, arXiv:1404.2188.[Online]. Available: http://arxiv.org/abs/1404.2188

[22] C. N. Dos Santos and M. Gatti, “Deep convolutional neural networksfor sentiment analysis of short texts,” in Proc. COLING, 2014,pp. 69–78.

[23] D. Zeng et al., “Relation classification via convolutional deep neuralnetwork,” in Proc. COLING, 2014, pp. 2335–2344.

[24] M. Kawato, K. Furukawa, and R. Suzuki, “A hierarchical neuralnetwork model for control and learning of voluntary movement,” Biol. Cybern., vol. 57, no. 3, pp. 169–185, Oct. 1987.

[25] C. Goller and A. Kuchler, “Learning task-dependent distributed representations by backpropagation through structure,” in Proc. IEEE Int. Conf Neural Netw., vol. 1, Jun. 1996, pp. 347–352.

[26] R. Socher, E. Huang, J. Pennin, C. Manning, and A. Ng, “Dynamic pooling and unfolding recursive autoencoders for paraphrase detection,” in Proc. NIPS, 2011, pp. 801–809.

[27] J. L. Elman, “Finding structure in time,” Cognit. Sci., vol. 14, no. 2, pp. 179–211, 1990.

[28] L. Fausett, Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.

[29] T. Mikolov, M. Karafiát, L. Burget, J. ˇ Cernock`y, and S. Khudanpur, “Recurrent neural network based language model,” in Proc. 11th Annu. Conf. Int. Speech Commun. Assoc., vol. 2, 2010, p. 3.

[30] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, “Extensions of recurrent neural network language model,” in Proc. IEEE ICASSP, May 2011, pp. 5528–5531.

[31] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. ˇ Cernock`y, “Strategies for training large scale neural network language models,” in Proc. IEEE Workshop Autom. Speech Recognit. Understand. Dec. 2011, pp. 196–201.

[32] J. Schmidhuber, “Learning complex, extended sequences using theprinciple of history compression,” Neural Comput., vol. 4, no. 2,pp. 234–242,Mar. 1992.

[33] S. El Hihi and Y. Bengio, “Hierarchical recurrent neural networks forlong-term dependencies,” in Proc. NIPS, 1996, pp. 493–499.

[34] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” NeuralComput., vol. 9, no. 8, pp. 1735–1780, 1997.

[35] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, andJ. Schmidhuber, “LSTM: A search space odyssey,” IEEE Trans. NeuralNetw. Learn.Syst., vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[36] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio,“On the properties of neural machine translation: Encoder-decoderapproaches,” 2014, arXiv:1409.1259. [Online]. Available: http://arxiv.org/abs/1409.1259

[37] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluationof gated recurrent neural networks on sequence modeling,” 2014,arXiv:1412.3555. [Online]. Available: http://arxiv.org/abs/1412.3555

[38] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translationby jointly learning to align and translate,” 2014, arXiv:1409.0473.[Online]. Available: http://arxiv.org/abs/1409.0473

[39] A. M. Rush, S. Chopra, and J. Weston, “A neural attention modelfor abstractive sentence summarization,” 2015, arXiv:1509.00685.[Online]. Available: http://arxiv.org/abs/1509.00685

[40] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model forabstractive summarization,” 2017, arXiv:1705.04304. [Online]. Available:http://arxiv.org/abs/1705.04304

[41] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, “Gated selfmatchingnetworks for reading comprehension and question answering,”in Proc. ACL, vol. 1, 2017, pp. 189–198.

[42] A. Vaswani et al., “Attention is all you need,” in Proc. NIPS, 2017,pp. 6000–6010.

[43] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencieswith gradient descent is difficult,” IEEE Trans. Neural Netw.,vol. 5, no. 2, pp. 157–166, Mar. 1994.

[44] V. Nair and G. E. Hinton, “Rectified linear units improve restrictedBoltzmann machines,” in Proc. ICML, 2010, pp. 807–814.

[45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for imagerecognition,” in Proc. IEEE CVPR, Jun. 2016, pp. 770–778.

[46] R. Kumar Srivastava, K. Greff, and J. Schmidhuber, “Highwaynetworks,” 2015, arXiv:1505.00387. [Online]. Available: http://arxiv.org/abs/1505.00387

[47] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Denselyconnected convolutional networks,” in Proc. IEEE CVPR, Jul. 2017,vol. 1, no. 2, pp. 4700–4708.

[48] D. Jurafsky and J. Martin, Speech & Language Processing. London,U.K.: Pearson Education, 2000.

[49] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilisticlanguage model,” J. Mach. Learn. Res., vol. 3, pp. 1137–1155,Feb. 2003.

[50] W. De Mulder, S. Bethard, and M.-F. Moens, “A survey on the applicationof recurrent neural networks to statistical language modeling,”Comput. Speech Lang., vol. 30, no. 1, pp. 61–98, Mar. 2015.

[51] R. Iyer, M. Ostendorf, and M. Meteer, “Analyzing and predictinglanguage model improvements,” in Proc. IEEE Workshop Autom.Speech Recognit. Understand., Dec. 1997, pp. 254–261.

[52] S. F. Chen, D. Beeferman, and R. Rosenfeld, “Evaluation metricsfor language models,” School Comput. Sci., Carnegie Mellon Univ.,Pittsburgh, PA, USA, Tech. Rep., Jan. 2008. [Online]. Available:https://kilthub.cmu.edu/articles/Evaluation_Metrics_For_Language_Models/6605324, doi: 0.1184/R1/6605324.v1.

[53] P. Clarkson and T. Robinson, “Improved language modelling throughbetter language model evaluation measures,” Comput. Speech Lang.,vol. 15, no. 1, pp. 39–53, Jan. 2001.

[54] M. Marcus, M. A. Marcinkiewicz, and B. Santorini, “Building a largeannotated corpus of English: The Penn Treebank,” Comput. Linguistics,vol. 19, no. 2, pp. 313–330, 1993.

[55] C. Chelba et al., “One billion word benchmark for measuring progressin statistical language modeling,” 2013, arXiv:1312.3005. [Online].Available: http://arxiv.org/abs/1312.3005

[56] M. Daniluk, T. Rocktäschel, J. Welbl, and S. Riedel, “Frustratinglyshort attention spans in neural language modeling,” 2017,arXiv:1702.04521. Online]. Available: http://arxiv.org/abs/1702.04521

[57] G. Hinton et al., “Deep neural networks for acoustic modeling in speechrecognition: The shared views of four research groups,” IEEE SignalProcess. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[58] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition withdeep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust.,Speech Signal Process.*, May 2013, pp. 6645–6649.

[59] T. Kenter, A. Borisov, C. Van Gysel, M. Dehghani, M. de Rijke,and B. Mitra, “Neural networks for information retrieval,” in*Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017,pp. 1403–1406.

[60] B. Hu, Z. Lu, H. Li, and Q. Chen, “Convolutional neural networkarchitectures for matching natural language sentences,” in *Proc. NIPS*,2014, pp. 2042–2050.

[61] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learningdeep structured semantic models for Web search using clickthroughdata," in *Proc. ACM CIKM*, 2013, pp. 2333–2338.

[62] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semanticrepresentations using convolutional neural networks for Web search,"in *Proc. 23rd Int. Conf. World Wide Web (WWW Companion)*, 2014,pp. 373–374.

[63] Z. Lu and H. Li, "A deep architecture for matching short texts," in*Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1367–1375.

[64] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matchingas image recognition," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016,pp. 2793–2799.

[65] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matchingmodel for ad-hoc retrieval," in *Proc. 25th ACM Int. Conf. Inf. Knowl.Manage. (CIKM)*, 2016, pp. 55–64.