What's Wrong with Deep Learning for Meaning Understanding

Rodolfo Delmonte[1]

[1]*Ca Foscari University of Venice, Italy*

Abstract

In this paper I will tackle the debated question of whether Deep Neural Networks "understand" Natural Language when they process it either for classification or for generation tasks. I will start by quoting work currently carried out in the field of explainable AI then I will look into the internal procedures adopted by current DNNs to cope with the enormous and untreatable dimension of vocabulary size due to the presence of rare words and the use of subwords or n-gram character sequences. To explain the inability of DNNs to reuse previously discovered knowledge I will deal with the issue of generalization and the lack of systematicity in learning models. I will also propose reasons to motivate the difficulty of generalizing with morphology-rich and weakly configurational languages like Italian. In particular vocabulary size and rare words need a different approach from the one used with English texts. Eventually I will propose some alternative ways to deal with the problems discussed in the paper.

Keywords

Deep Learning, Meaning Understanding, Explainability, Adversarial Datasets, Word Embeddings and OOV Words

1. Introduction

In this paper I will not produce the nth experiment with Deep Neural Networks (hence DNNs) in order to show that modifying the learning rate or some other hyperparameter it is possible to achieve a slight increase in accuracy or F1 for a given collection of datasets. I will rather try to prove that the idea of increasing the size of training data to a giant amount will in no way help researchers to discover the way in which humans use language to communicate. I will criticize the way in which predictions are generated by word embeddings after having pruned or dropped the long tail of rare words from the vocabulary frequency list, substituting it with meaningless subword units which have no context. I will also criticize the use of leaderboards like GLUE [1] to assert that DNNs are capable to understand language because they can cope with tasks such as NLI or TextEntailment with remarkable accuracy or F1 score. Ultimately, I will list a series of alternative approaches which reject the idea that unsupervised and/or self-learning models constitute the only way to carry out fruitful advances in the field of language understanding. In sum, the paper constitutes a deep reflection on state of the art of DNNs for language understanding.

Current DNNs are data hungry, even though the late appearance of pre-trained models have changed the experimental setup by facilitating their use. This notwithstanding, every new reference model is still data-hungry and they have been so for the last 40 years or so when in 1989 the whole story started. It happened in the conference NATO-ASI held in Cetraro(Italy) where the team from IBM presented for the first time the language model for speech recognition. At the end of the social dinner, after having drunk a sizeable amount of alcohol, everybody decided we had to sing a song while doing the conga line around the tables, and the words were the 5 following ones: "We need more training data".

In this paper I will focus on the fundamental issues regarding Deep Learning (hence DL) and Language Understanding, i.e. whether or not – in the case of NLP - it is aiming at natural language understanding as their supporters maintain. For a start, I will assume that Deep Learning is a highly sophisticated and very powerful technique for (co-occurrence) pattern matching whose tremendous success is in line with the achievements of ASR (Automatic Speech Recognition). Both NLP and ASR are now best processed by Transformers [2] and make use of the self-attention mechanisms which allow training to start directly from raw or unsupervised data, that in the case of NLP corresponds to raw (curated) text.

In fact, what DNNs are really doing is predicting sequences of words or sentences without any associated meaning. On the contrary, among supporters and inventors of this techniques, there is the belief that DNN understand (the meaning of) natural language they are predicting. Geoff Hinton[1] and Yann LeCun[2], the two most prominent researchers in the field of Deep Learning and NLP, have many times declared that neural networks are a computational embodiment of the way in which the human brain learns language. Yoshua Bengio in a recent article[5], declared that current machine learning methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice.

As for computer video, Yann LeCun[6] describes the situation as more dramatic, seen that when using knowledge stored in pretrained models of video images difficulties arise when trying to recognized a new unseen set of frames. Representing missing video frames or patches is almost impossible due to the infinite number of video frames that can be added.[3]

However, we don't intend to dismiss completely DNN and think they are useful tools but we need a better way to do that. The question I will pose at the start is "Why is it that Deep Neural Networks (DNN) make such terrible mistakes when systems are regularly deployed?", and what may be the causes for such error prone behaviour. At the end of the paper I will comment on a number of proposals for a more convenient and promising way of using DL in language related tasks.

---

[1]He said: "I think that the most exciting areas over the next years will be really understanding text and videos. I will be disappointed if in five years' time we do not have something that can watch a YouTube video and tell a story about what happened."[3]

[2]said "The next big step for Deep Learning is natural language understanding, which aims to give machines the power to understand not just individual words but entire sentences and paragraphs."[4]

[3]In their words "There are an infinite number of possible video frames that can plausibly follow a given video clip. It is not possible to explicitly represent all the possible video frames and associate a prediction score to them. In fact, we may never have techniques to represent suitable probability distributions over high-dimensional continuous spaces, such as the set of all possible video frames."

2. Exploring and Explaining Weaknesses and Limitations of DLM

I will start by discussing what's wrong with the use of DNNs by commenting a recent paper by Camburu et al.[7], where the aim of the researchers was showing whether current post-hoc explanatory methods[4] are able to find out if DNNs achieve trustable and reliable results. These methods have been developed with the goal of shedding light on highly accurate, yet black-box machine learning models. In fact their results are not clarifying what is going on in the hidden layers of the DNN. On the contrary, the explainers predict the most relevant token to be among the tokens with zero contribution. When applied to real-world neural networks, explainers are commonly validated under the assumption that the learned models behave reasonably. However, neural networks often rely on unreasonable correlations, even when producing correct decisions. NN rely more often on spurious and irrelevant correlations as discussed below. The most common instance-wise explanatory methods are feature-based, i.e., they explain a prediction in terms of the input unit-features (e.g., tokens for text and super-pixels for images). There are two major types of explanations: (i) feature-additive: provide signed weights for each input feature, proportional to the contributions of the features to the model's prediction; (ii) feature-selective: provide a (potentially ranked) subset of features responsible for the prediction; (iii) example-based: identify the most relevant instances in the training set that influenced the model's prediction on the current input. In their paper they tested three among the most common explainers and found out that all explainers are prone to stating that the most relevant feature is a token with zero contribution.

Before dealing with the lack of generalization or failures of transportability I will rather comment on two events that happened lately (September 1st), related to terrible mistakes made by Facebook. Facebook's policy on the use of AI algorithms is guided by Yann Lecun, a man who strongly supports the use of unsupervised models or rather self-supervised models as they are now called. The decision whether contents to appear on FB contain dangerous and derogatory material for minorities or for the general public is eventually left to be made by the algorithm. Errors will be backed by backpropagation in the network. When the network chooses the wrong candidate the weights are lowered at all edges that enter that node. In this way the wrong association between a given training example and the current observation will not be properly weighted. Current ML or DL algorithms are called "Frankenalgorithms" menacing our lives because they are incomprehensible.[8])

The algorithm that has the task to filter out unwanted public material erased the pages of anticomplotists' Wu Ming 1 (alias Roberto Bui the author of La Q di Qomplotto (edizione Alegre)) and a number of similar social networks (Magazzino Parallelo and ARCI Cesena) who were working on a public document on the same topic, because of its inability to tell negation(anti) of negative values (complotists like QAnon). This is the "double negation" curse which is one of the hardest task in language processing. More recently September 23rd, all social networks on Facebooks related to "Stazione dell'Arte"/Art Station organizing the largest exhibition of the Sardinian artist Mirella Bentivoglio have been cancelled. Also the page showing the art work "Fiore Nero"/Black Flower has been obscured reason being it incites to hate. The art work

---

[4]They test three most popular explanatory methods with the corresponding explainer algorithm, and they are: LIME, SHAP and L2X.

denounces discrimitative behaviours: what must have gone wrong is the representative image – which is a patchwork of newspaper articles dealing with black dresses, black horses, black flowers and black coffins, is interpreted again as a negative image which may incite violence. Deep Learning(hence DL) is regarded a success in classifying written texts, however as shown by the two examples above, negation and negative events are not easy to disentangle and may totally confound DL. But that is not the only failure of DL.[5]

2.1. Biases, Generalization and Transportability

In their paper Bender et al.[9], come to the following conclusion about the possibility of a LM to approximate language understanding: "a stochastic parrot". If DL could really learn language like children do it in their process of language acquisition it would be possible to benefit from **Systematicity** while coping with **Variability**. Systematicity means that children show a remarkable ability to infer the phonological, structural, lexical and semantic system of language. However, none of this kind happens in DL because neural networks are unable to produce inferences from what they previously learnt, and apply it to new unseen text, as clearly shown in[10]. Ruis et al.[11] comment on the ability humans have to generalize the use of new words learnt into a variety of contexts, and this ability is ascribed to our "aptness for systematic compositionality". This aptness is then linked to our "algebraic" capacity to understand and produce potentially infinite linguistic combinations at sentence or phrasal level from known linguistic components. The same authors blame modern deep neural networks, which have shown their powerful capacity in many domains – in particular in challenging tasks such as NLI and Text Entailment -, but "have not mastered comparable language-based generalization challenges", and this is clearly a fact that underlies "their sample inefficiency and inflexibility". Systematic, rule-based generalization is at the core of the recently introduced SCAN dataset. In a series of studies, Lake et al.[12], tested various standard deep architectures for their ability to extract general compositional rules supporting zero-shot interpretation of new composite linguistic expressions. In most cases, as reported by the authors, neural networks were unable to generalize correctly. These results demonstrate the challenges of accounting for common natural language generalization phenomena with standard neural models.

In order to show that DNNs "understand language" a public leaderboard has been organized called GLUE[1] – General Language Understanding Evaluation benchmark, followed by SuperGLUE lately - containing eight language understanding tasks. The tasks can be regarded as benchmarks against which to measure the ability of the system to cope with specific issues in language understanding, and are all made up by existing data, accompanied by a single-number performance metric and an analysis toolkit. However, as has been argued in a number of papers, these tasks contain surprising cues that allow DNNs to perform better. In their paper on SuperGLUE, Wang et al.[13] comment on the continual improvements achieved and surpassing non-expert human performance. As they mention, progress has been made possible by increasing computer power, data quantity and as a consequence model increasing capacity.

---

[5]On the web it is possible to find a host of example cases of DL mistakes that can be accessed – I have found two websites with a long list of AI failure stories: https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020https://medium.com/money-talks-the-official-abe-blog/how-to-fail-with-artificial-intelligence-b3c4b1966bb3

But also improvements on the algorithm producing the model decoding plays an important role, basically in the ability to encode more and more information on the context surrounding each token. However, as the authors have to admit, "While some initially difficult categories saw gains from advances on GLUE (e.g., double negation) , others remain hard (restrictivity) or even adversarial (disjunction, downward monotonicity)… it remains difficult to extract many details crucial to semantics without the right kind of supervision". In other words, fully unsupervised approaches are non appropriate with heavily semantically based tasks.

The ability to generalize to new and unseen cases is the inference predicting power of the model. In BERT, the inference ability is created by the MLM (Masked Language Model) which is taught to predict one or more masked token/s in the training corpus – the cloze task. However, the algorithm is prevented from taking vector values associated to the actual masked token and forced to choose the ones of the closest token, in order to avoid overfitting. Overfitting and the lack of ability to generalize to new instances is however the real missing point. This may be due to the so-called "sampling bias", i.e. the presence of noise, a lot of errors in the enormous amount of data used to produce the model. But - in the case of transfer learning with BERT[14] - it may also be related to the choice of a surrogate token and its word embeddings, which may be close to the original token and may have been chosen only by chance. This introduces an anomaly in the computation that doesn't harm accuracy in case the test set is taken from the same corpus of the training set. More concretely, Le Bras et al.[15] raise the question of whether DNN models have learned to solve a dataset rather than the underlying task by overfitting to spurious dataset biases. For instance, the success in the use of Universal Sentence Embedders (hence USE) by most DNNs in heavily semantic downstream tasks may be due to superficial heuristics (as supposed in[16]) for the NLI - Natural Language Inference) and not to a deep modeling of semantic features.

The existence of such biases is argumented in[15] where they describe an experiment in which biases are filtered out. Their approach is called AFLITE, and it is meant to adversarially filter dataset biases, as a means to mitigate the prevalent overestimation of machine performance. The results show a better generalization ability with out-of-distribution task but also a dramatic drop in performance from 92% to 62% accuracy for a heavily semantically oriented dataset, SNLI. The same happens with another semantically heavy datase, the ARCT (Argument Reasoning Comprehension Task), as reported by Niven et al.[17] in a paper where they analyse the results obtained by BERT. They discovered that what triggers "comprehension" and consequent decision-making in the correct choice is presence of particular non relevant cues, like "not" and other spurious or incoherent statistical cues . After the discovery, the authors modified the test set accordingly eliminating the special cues and ran the experiment with BERT achieving just 53% accuracy, against a previous 77%.

Now let's consider what happens when a new domain or a new genre is chosen. In their paper, Marshall et al.[16] focus on the lack of transportability of pretrained models in a new context, a fact already shown by[18]: permutations of training data implies substantial changes in performance. In their introduction, Marshall et al. note that the lack of transportability – without retraining - for NLP tasks "has been raised by healthcare experts who have expressed their frustration in the limitations of algorithms built in research settings for practical applications and the reduction of performance outside of their development frame". More generally, Pearl[19] remarks that "current systems lack the ability to recognize or react to new circumstances they

have not been specifically programmed or trained for". Word embeddings' accuracy degrades when used over different datasets according to[20]. In other words, the universal model does not exist yet that can be used for any new dataset for the same task for which it was created. Experiments carried out by Marshall et al. show the degradation of performance when moving models created for the NLI from one dataset to another, using SciTail, MultiNLI and SNLI datasets. When training and testing is done in the same dataset accuracy score vary from 93.08 to 83.5 for MultiNLI and 90.4 for SNLI; but when datasets are changed accuracy drops down to 52.72 for SciTail with SNLI test set, 44.49 for MultiNLI with SciTail, and 44.2 for SNLI with SciTail. Better results are obtained when using MultiNLI as variation dataset. Similar degradation is observed for NER (Name Entity Recognition) task. In a previous paper on explainability for MRC (Machine Reading Comprehension), another task which is based on language understanding and reasoning, researchers from the same Department[21] focus on the ability of models to develop explanations which go beyond the surface form of the text on which they have been trained with similar results.

There is an important component of this kind of research which has gone unnoticed so far, and it is the fact that GLUE and other similar benchmarks are all wrought upon the English language. English is a strongly configurational language that preserves the linear structure of its syntactic component, thus failing to constitute a valid example for weakly configurational languages like Chinese, Japanese, but also Russian, Serbian, Italian. In particular, in a paper on lexical and syntactic complexity measures, Delmonte[22] compares the non-projectivity parameter associated to Italian (6.65%) as being very close to Latin (7%) versus American English as attested by PTB (Penn TreeBank) which is well below 1%. In particular, positional encoding which is used in transformer architecture to enforce contextual dependencies, in fact reduces its usefulness in texts belonging to weakly configurational languages.

3. Word Embeddings and DSM

Most ML and DL systems are based on Word Embeddings which in turn are theoretically based on DSM – Distributed Semantic Models. According to the theory, input word embeddings should contain a certain amount of syntactic and semantic information that is then used by each layer of the neural network to find the most similar vector in the pre-trained model in order to make a classification choice or receive a label. In fact, word embeddings may contain morphological variations, but also homographs non homophones (récord/recòrd), in case of polysemous words (and most frequent words usually have a long list of homographs with different meanings) they will all be packed in the same embedding but with different context (hopefully). According to experiments carried out by[23] using BERT, the eight nearest (measured by cosine similarity) neighbours of the word DOG are "dogs (0.67), cat (0.44), horse (0.42), animal (0.38), canine (0.37), pig (0.37), puppy (0.37), bulldog (0.37), and hound (0.35)", but we don't know in which way this happened and it is certainly most akward and unpredictable for humans to make such associations. Even though all eight candidates share the same semantic domain, we may wonder why "horse" gets better score than say "canine"; the same applies to "pig" which gets a better score than "puppy" or "bulldog".

The use of word embeddings is now taken for granted[6] as the best way to encapsulate syntax and semantics related to a given word. Embeddings may be considered from two connected points of view: the presence of OutOfVocabulary words and the way in which similar candidates are chosen by means of distance measure related to spatial dimensions like sine/cosine.

Let's consider how Out of Vocabulary Words are treated in BERT, where subword vectors are used for representing both the input text and the output tokens. When an unseen word is presented to BERT, it will be sliced into multiple subwords, even reaching character subwords if needed. That is how it deals with unseen words. In Ruzzetti et al.[25] a new method is proposed to build embeddings of OOV words starting from dictionary definitions, which are taken from WordNet. Subword tokenization algorithms rely on the principle that frequently used words should not be split into smaller subwords, but rare words should be decomposed into meaningful subwords. For instance "inapprehensible" might be considered a rare word and could be decomposed into "in" "ap" "pre" "hen" and "s" "ible". Both "pre" and "hen" as stand-alone subwords would appear more frequently while at the same time the meaning of "inapprehensible" is lost by the composite meaning of "pre" and "hen". In other word, in some cases it is possible for a OOV word to be represented as, at the very least, the collection of its meaningless bigrams or trigrams and in all these case subword and character tokens will be used to generate embeddings for, thus no longer retaining the composite meaning. Subword embedding vectors will be then averaged to generate an approximate vector for the original word.

A better way to deal with n-gram subwords is proposed in[26], where a system for subword-based compositional word embedding is presented. The model is called probabilistic bag-of-subwords (PBoS), as it applies bag-of-subwords for all possible segmentations based on their likelihood. The hidden assumption here is that words are made of meaningful parts (cf. morphemes) and that the meaning of a word is related to the meaning of their parts. As we saw previously, a word embedding vector is composed by taking the sum or average of the vectors of the subwords (character n-grams) that appear in the given word. However, the importance of different subwords is ignored since all of them are given the same weight. In the PBoS, the subword segmentation part is a probabilistic model capable of handling ambiguity of subword boundaries and ranking possible segmentations based on their overall likelihood. The final word embedding vector is then the probabilistic expectation of all the segmentation vectors. This blurs the boundary between words and non-words, and automatically enables the system to handle unseen words, alternative spellings, typos, and nonce words as normal cases. This approach is similar to the one used for languages using ideographs like Chinese, Korean and Japanese (see[27]). In their experiments, presence of OOV in the corpus has an adversarial impact on task performance. Mitigation methods for the presence of OOV words are applied based on subwords with their context which they try to cover in a preliminary phase when analysing the task corpus, prioritizing frequent OOV subwords. Languages with a rich

---

[6]In Devlin[14] we find this statement: "I'm not sure what these vectors are, since BERT does not generate meaningful sentence vectors. It seems that this is doing average pooling over the word tokens to get a sentence vector, but we never suggested that this will generate meaningful sentence representations." In fact, in a later paper by Reimers and Gurevych[24] a solution has been suggested to encapsulate a complete sentence embeddings into a single vector representation and use cosine measures to compare sentence pairs. However this procedure will not produce semantically valuable representations.

morphology on the contrary, require the long tail of rare words to be preserved as discussed in a paper[28] on Italian word embeddings. They discover that the "accuracy increases as the number of dimensions of the embedded vectors increases. This indicates that Italian language benefits of a rich representation that can account for its rich morphology." They tested an extended number of hyperparameters on the Analogy dataset and found that an increased value of vector dimension and in the number of negative samples gave best results. This meant taking into account also very infrequent words and a much larger vocabulary.

4. Producing Meaningful Communication

Despite enormous improvements in hardware and data collection have lead to achieve impressive performance in language modeling and a number of NLP related classical tasks, what is missing from these approaches is the ability to produce meaningful communication. The lacking feature seems to be commonsense ability that characterize humans in their behaviour. According to Yann Lecun[6], common sense could be achieved by machines using AI techniques like self-supervised learning[7] However this is not what experimental work tells us. In a paper by Bisk et al.[29] language understanding is contextualized into linguistic communication based on a shared experience of the world. "Meaning does not arise from the statistical distribution of words, but from their use by people to communicate. Many of the assumptions and understandings on which communication relies lie outside of text"[29]. Modern approaches to learning dense representations allow us to better estimate these distributions from massive corpora. However, modeling lexical co-occurrence[8], no matter the scale, is still the way in which modeling the written world takes place. Models constructed this way blindly search for symbolic co-occurences void of meaning. Chevalier-Boisvert et al.[30] formulate the following question in their Introduction "How can a human train an intelligent agent to understand natural language instructions?". In their paper, they present the BabyAI research platform, whose purpose is to facilitate research on grounded language learning and express the following final evaluation: "We put forward strong evidence that current deep learning methods are not yet sufficiently sample-efficient in the context of learning a language with compositional properties." Past work has found that pretrained word and sentence representations fail to capture many grounded features of words[31] and sentences, and current NLU systems fail on the thick tail of experience-informed inferences, such as hard coreference problems[32].[9]

5. Alternative Approaches

In this section I will present a number of alternative approaches to the "unsupervised" universal use of DNNs for natural language understanding purposes. Nye et al.[34] present a neuro-

---

[7]In Yann Lecun words: "We believe that self-supervised learning (SSL) is one of the most promising ways to build such background knowledge and approximate a form of common sense in AI systems." And further on "self-supervised learning may be helpful in unlocking the dark matter of intelligence"

[8]remember that the fundamental elements constituting all the computation in DNN are word co-occurence frequency absolute/relative value and number of pixel

[9]The following test sentence "I parked my car in the compact parking space because it looked (big/small) enough." still presents problems for text-only learners[33]

symbolic model which learns entire rule systems from a small set of examples, which, rather than predicting directly outputs from inputs, pass through an intermediate explicit set of rules. The approach involves using program synthesis to learn explicit rule systems from just a few examples. Compared to pure neural approaches, the approach has two main advantages: robustness and interpretability. They train the model to induce the explicit system of rules governing a small set of previously seen examples, drawing upon techniques from the neural program synthesis literature. As the authors specify, "the approach combines a neural "proposer" and a symbolic "checker". The system output is fully explainable in this manner because the representation produced by the model is a symbolic program. It is also interpretable: when mistakes are made, the incorrect program can be analyzed in order to understand the error.

Lake et al.[12] criticize the idea of using seq2seq models to cope with problems that require compositional generalization and the ability to do that in new domains and unseen text. Then they describe in detail a solution based on what they call meta-seq2seq framework, which requires meta-training and meta-learning. This requires memory-augmented neural networks and a different architecture and the aim is to allow the network to acquire the compositional skill needed to solve new problems. New seq2seq problems are solved entirely using the activation dynamics and external memory of the networks; no weight updates are made after the meta-training phase ceases.[10]

Gary Marcus[35, 36, 37] focuses on the need to introduce "reasoning" into the learning process and defines a program to do that. It is a four-step program: initial development of hybrid neuro-symbolic architectures, followed by construction of rich, partly-innate cognitive frameworks and large-scale knowledge databases, followed by further development of tools for abstract reasoning over such frameworks, and, ultimately, more sophisticated mechanisms for the representation and induction of cognitive models. According to Marcus, progress towards these four prerequisites could provide a substrate for richer, more intelligent systems than are currently possible. Learning[11] the meaning or the content of written text is still a challenge for DL, even if the size of the training corpus has increased to an incredible amount. DL is unable to learn abstract concepts from examples as would human beings. Always according to Marcus, these problems "include the need to bring meaning and reasoning into systems that perform natural language processing, the need to infer and represent causality, the need to develop computationally-tractable representations of uncertainty and the need to develop systems that formulate and pursue long-term goals."

Another important and promising substitute for the DNNs is Quantum Mechanics which

---

[10]This is how they describe the new framework:"As is standard with meta learning, training is distributed across a series of small datasets called "episodes" instead of a single static dataset, in a process called "meta-training." Specific to meta seq2seq learning, each episode is a novel seq2seq problem that provides "support" sequence pairs (input and output) and "query" sequences (input only). The network loads the support sequence pairs into an external memory to provide needed context for producing the right output sequence for each query sequence. The network's output sequences are compared to the targets, demonstrating how to generalize compositionally from the support items to the query items." And further on: "Through its unique choice of architecture and training procedure, the network can implicitly learn rules that operate on variables, an ability considered beyond the reach of eliminative connectionist networks but which has been pursued by more structured alternatives."

[11]It is also the same concept of learning that has to be reconsidered, "leading to a (perhaps new) form of learning that traffics in abstract, language-like generalizations, from data, relative to knowledge and cognitive models, incorporating reasoning as part of the learning process".

is used for expressing adequately Quantum Logics . Sordoni and colleagues[38] developed a Quantum Language Model (QLM), a generalization of prior language modeling approaches that provides the means to model term dependencies without severing the connection between the probability of observing a multiword expression and the probabilities of observing its component terms. van Rijsbergen[39] and Melucci[40] are responsible for explaining how tensor product representations and quantum entanglement can be represented with subspaces, and this affects the way that categories or natural kinds might be modelled in an IR system. Details of this proposal have been put forward by an extended number of people, besides the ones cited above, including[41, 38, 42]. The example used is "ivory tower", whose meaning is not a linear composition of the individual meanings of "ivory" and "tower", instead it carries a new meaning. Thus a new language modeling paradigm is proposed based on Quantum Probability to account for such intricate non-linear combination of word meanings. The model proposed is the Semantic Hilbert Space which is used to formulate quantum-like phenomena in language understanding, and to model different levels of semantic units in a unified space[43] . This is achieved through an end-to-end neural network architecture, which provides means for training the network components. Each component corresponds to a physical meaning of quantum probability with well-defined mathematical constraints. Moreover, each component is easier to understand than the kernels in convolutional neural network and cells in recurrent neural networks. Thus the main outstanding feature of the Quantum Probability driven network is the nature of their components, i.e. "Self-Explainability". It is a bottom-up architecture that represents each level of semantic units in a uniform SHS, from the sememe to sentence representation. The framework is made operative through "superposition, mixture and semantic measurement". As the authors comment on the related experiment made on six benchmarking text datasets, the performance is effective, robust and most of all "self-explainable".

Finally, we include the point of view expressed in the paper by Bender et al.[44] on the environmental impact of DNNs and on the need to abandon the search for larger and larger datasets without curating them appropriately or investigating on the existence of derogatory language that might offend social groups. This is how they express these concepts: "This means making time in the research process for considering environmental impacts, for doing careful data curation and documentation, for engaging with stakeholders early in the design process, for exploring multiple possible paths towards long-term goals, for keeping alert to dual-use scenarios, and finally for allocating research effort to harm mitigation in such cases."

6. Conclusion

In this paper we have assumed that DNNs are powerful classifiers but have poor predictive power because of their inability to reuse the knowledge encoded in their model for new domains. In particular we have shown that DNNs lack the essential quality called Systematicity, that is the ability to apply learned relations to new and unseen contexts in a fully trustable and reliable manner; the need to reduce dimensions of unique wordforms vocabulary has introduced the technique of word decomposition on a bag-of-subword level, thus destroying any possible meaning-wordform relation; subword vocabularies have been used to cope with less frequent wordforms with the excuse that that would mitigate the impact of OOV words but this has

negative impact in languages different from English, morphological rich and also scripta languages which are sign rich; transfer learning is a modality that does away with the masked word used by Transformers to be predicted, rather it accepts a token whose word embedding vector cosine value approximates it thus introducing an anomaly into the backpropagation process; the gradient descent score is computed by a negative cross entropy log-probability loss formula minimizing the (mean square) error and trying to converge to an optimum, which is in fact favouring the least different approximation rather than the most similar one[12]; all semantically heavy benchmarks are in English, a language that does not represent - from a syntactic structural point of view - the majority of mostly used languages in the world. Rather English constitutes a statistically valid baseline being strongly configurational and thus structurally strongly invariant; eventually, knowing distributional co-occurence features of a word/phrase/paragraph or sentence does in no case correspond to knowing its meaning. Understanding language requires the ability to make inference, causality and reasoning with uncertainty, abilities totally lacking in current DNNs. To summarize our position towards the use of LMs and DNNs we assume Bender et al.'s argument in favour of a more ethical and sustainable research framework. In particular we fully subscribe to the need to weigh the risk of "substantial harms, including stereotyping, denigration, increases in extremist ideology, and wrongful arrest" represented by gigantic pretrained models. NLP researchers should investigate the possibility of utilizing other techniques in approaching NLP tasks that are effective without being endlessly data hungry and harmful - even though the use of pre-trained models has now reduced the need to compute a new model for every experiment.

Acknowledgments

References

[1] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Glue - a multi-task benchmark and analysis platform for natural language understanding, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=rJ4km2R5t7.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 6000−6010.

[3] G. Hinton, Machine learning, 2017. URL: https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama-geoffrey-hinton.

---

[12]In fact this procedure is used in all other statistically based tasks like ASR - word error rate - and machine translation and aims at reducing the loss on unseen data

[4] Y. LeCun, Y. Bengio, Self-supervised learning is the key to human-level intelligence, 2021. URL: https://venturebeat.com/2020/05/02/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/.

[5] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, C. Pal, A meta-transfer objective for learning to disentangle causal mechanisms, in: Proceedings of 8th International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/pdf?id=ryxWIgBFPS.

[6] Y. LeCun, I. Misra:, Self-supervised learning - the dark matter of intelligence, 2021. URL: https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/.

[7] O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz1, P. Blunsom, Can i trust the explainer? verifying post-hoc explanatory methods, preprint, 2019.

[8] Y. Wilks, Artificial Intelligence – Modern Magic or Dangerous Future, Icon Books, 2019.

[9] T. G. S. S. Emily M. Bender, Angelina McMillan-Major, On the dangers of stochastic parrots - can language models be too big?, in: Proceedings FAccT '21, ACM, 2021, pp. 610–623.

[10] H. Yanaka, K. Mineshima, D. Bekki, K. Inui, Do neural models learn systematicity of monotonicity inference in natural language?, 2020.

[11] L. Ruis, J. Andreas, M. Baroni, D. Bouchacourt, B. M. Lake, A benchmark for systematic generalization in grounded language understanding, preprint, 2020.

[12] B. M. Lake, Compositional generalization through meta sequence-to-sequence learning, preprint, 2019.

[13] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, preprint, 2019.

[14] J. Devlin, M. Chang, K. Lee, K. Toutanova, Bert - pre-training of deep bidirectional transformers for language understanding, in: Proceedings NAACL: human language technologies, volume 1 (long and short papers), Association for Computational Linguistics, 2019, pp. 4171–4186.

[15] R. L. Bras, S. Swayamdipta, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal, Y. Choi, Adversarial filters of dataset biases, in: Proceedings of the 37th International Conference on Machine Learning, 2020.

[16] G. Marshall, M. Thayaparan, P. Osborne, A. Freitas, Switching contexts: Transportability measures for nlp, in: Proceedings of the 14th International Conference on Computational Semantics, ACL, 2021, pp. 1–10.

[17] T. Niven, H.-Y. Kao, Probing neural network comprehension of natural language arguments, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019, p. 4658–4664.

[18] N. Schluter, D. Varab, When data permutations are pathological - the case of neural natural language inference, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, ACL, 2018, pp. 4935–4939.

[19] J. Pearl, The seven tools of causal inference, with reflections on machine learning, Communications of the ACM 62 (2019) 54–60.

[20] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, H. Veisi, Sentiment analysis based on improved pre-trained word embeddings, Expert Systems with Applications (2019) 139–147.

[21] M. Thayaparan, M. Valentino, A. Freitas, A survey on explainability in machine reading

comprehension, preprint, 2020.

[22] R. Delmonte, Syntactic and lexical complexity in italian noncanonical structures, in: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, COLING, 2016, pp. 67–78.

[23] B. M. Lake, G. L. Murphy, Word meaning in minds and machines, preprint, 2021.

[24] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[25] E. S. Ruzzetti, L. Ranaldi, M. Mastromattei, F. Fallucchi, F. M. Zanzotto, Lacking the embedding of a word? look it up into a traditional dictionary, preprint, 2020.

[26] X. Z. Y. L. Zhao Jinman, Shawn Zhong, Pbos: Probabilistic bag-of-subwords for generalizing word embedding, in: Proceeding of EMNLP - Findings of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 596–611.

[27] N. O. Sangwhan Moon, Effects and mitigation of out-of-vocabulary in universal language models, Journal of Information Processing 29 (2021) 490–503.

[28] R. Tripodi, S. L. Pira, Analysis of italian word embeddings, preprint, 2017.

[29] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, J. Turian, Can i trust the explainer? verifying post-hoc explanatory methods, preprint, 2020.

[30] M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, Y. Bengio, Babyai - a platform to study the sample efficiency of grounded language learning, in: Proceedings ICLR, arXiv:1810.08272v4, 2019.

[31] L. Lucy, J. Gauthier, Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning, in: Proceedings of the First Workshop on Language Grounding for Robotics, Association for Computational Linguistics, 2017, pp. 76−−85.

[32] D. R. Haoruo Peng, Daniel Khashabi, Solving hard coreference problems, in: Proceedings of NAACL: Human Language Technologies, ACL, 2015, pp. 809–819.

[33] M. van Schijndel, A. Mueller, T. Linzen, Quantity doesn't buy quality syntax with neural language models, preprint, 2019.

[34] M. I. Nye, A. Solar-Lezama, J. B. Tenenbaum, B. M. Lake, Learning compositional rules via neural program synthesis, preprint, 2020.

[35] G. Marcus, Deep understanding - the next challenge for ai, in: Proceedings NeurIPS-2019, 2019.

[36] G. Marcus, GPT-2 and the nature of intelligence, The Gradient, 2020.

[37] G. Marcus, The next decade in ai, four steps towards robust artificial intelligence, preprint, 2020.

[38] A. Sordoni, J.-Y. Nie, Y. Bengio, Modeling term dependencies with quantum language models for ir, in: Proceedings of the 36th International ACM-SIGIR conference on research and development in Information Retrieval, ACM, 2013, p. 653—662.

[39] C. van Rijsbergen, The geometry of information retrieval, Cambridge University Press, Cambridge, UK, 2004.

[40] M. Melucci, Relevance feedback algorithms inspired by quantum detection, IEEE Transactions on Knowledge and Data Engineering 28 (2015) 1022−1034.

[41] P. Basile, A. Caputo, G. Semeraro, Encoding syntactic dependencies by vector permutation,

in: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 2011, pp. 43–51.

[42] B. Wang, Q. Li, M. Melucci, D. Song, Semantic hilbert space for text representation learning, presented at WWW '19, 2019.

[43] D. Widdows, T. Cohen, Reasoning with vectors - a continuous model for fast robust inference, Logic Journal of IGPL 23 (2015) 141–173.

[44] E. M. Bender, A. Koller, Climbing towards nlu - on meaning, form, and understanding in the age of data, in: Proceedings 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5185–5198.