Shades of Knowledge-Infused Learning for Enhancing Deep Learning

Amit Sheth, Manas Gaur, Ugur Kursuncu, and Ruwan Wickramarachchi
University of South Carolina

***Abstract*—Deep Learning has already proven to be the primary technique to address a number of problems. It holds further promise in solving more challenging problems if we can overcome obstacles, such as the lack of quality training data and poor interpretability. The exploitation of domain knowledge and application semantics can enhance existing deep learning methods by infusing relevant conceptual information into a statistical, data-driven computational approach. This will require resolving the impedance mismatch due to different representational forms and abstractions between symbolic and statistical AI techniques. In this article, we describe a continuum that comprises of three stages for infusion of knowledge into the machine/deep learning architectures. As this continuum progresses across these three stages, it starts with shallow infusion in the form of embeddings, and attention and knowledge-based constraints improve with a semideep infusion. Toward the end reflecting deeper incorporation of knowledge, we articulate the value of incorporating knowledge at different levels of abstractions in the latent layers of neural networks. While shallow infusion is well studied and semideep infusion is in progress, we consider *Deep Infusion of Knowledge* as a new paradigm that will significantly advance the capabilities and promises of deep learning.**

■ **For many, the** purpose of artificial intelligence (AI) has been to achieve human-level intelligence. In that direction, recent years have seen data-driven machine learning (ML) models, specifically neural networks, acquiring remarkable success in an increasing number of tasks, such as object detection in images and speech recognition. On the other hand, these approaches proved to be limited in their ability to perform the tasks with generality,

Digital Object Identifier 10.1109/MIC.2019.2960071
Date of current version 17 January 2020.

1089-7801 © 2019 IEEE Published by the IEEE Computer Society

adaptability, explainability, toward pursuing "machine intelligence." As the dependence over large datasets is critical, the challenge is more acute since there is a lack of adequate and high-quality labeled data. Moreover, such a dataset may not cover all possibilities concerning the task in question, including those likely to arise in the future. In natural language understanding (NLU), for example, algorithms have not yet progressed to capture the implicit contextual meaning of the content. One approach to address such limitations and make intrinsically more intelligent systems is to combine the bottom-up data-dependent processing with top-down processing, as observed by the cognitive scientists and to a lesser extent by computer scientists.[11]; Yang *et al.* 2017. The blending of deep/machine learning with structured knowledge (e.g., knowledge graphs), which we call "Knowledge-Infused Learning,"[7,8] is an approach to address challenges such as: first, decreasing the dependence on large datasets, second, reducing bias in the dataset, third, providing ability to trace information allowing explainability of a model, fourth, improving the search space for information specific to a domain since anomalies, irregularities and edge cases for which there may not be a large dataset to learn from, fifth, reducing the complexity of model architecture, and sixth, reducing false alarm in performance of a model. There have been early attempts at using external knowledge in machine learning to address these challenges; however, there is a long way to go to achieve true potential.

> "KGs will play an increasing role in developing hybrid neuro-symbolic systems (that is bottom-up deep learning with top-down symbolic computing) as well as in building explainable AI systems for which KGs will provide scaffolding for punctuating neural computing."

In the past decade, as symbolic or logical approaches to AI garnered substantial research attention, significant advances have come from statistical learning approaches. While these approaches were seen as complementary to each other,[10] their integration in one computational framework will be pivotal for pursuing machine intelligence with increased generality, adaptability, and explainability. This also has the potential of better supporting integrated top-down and bottom-up processing that the human brain appears to do well so seamlessly. Building upon the prior observations on the importance of knowledge in learning (e.g., data alone is not enough,[4] the machine will propel machine understanding,[11] we posit that knowledge, nowadays represented as knowledge graphs (KGs), will be the key enabler.

Learning the underlying patterns in the data goes beyond instance-based generalization to some external knowledge represented in the structured graphs or networks. The deep learning (DL) has shown significant advances in improving natural language processing (NLP) by probabilistically learning latent patterns in the data using a multilayered network of the computational nodes (i.e., neurons/hidden units). However, with the tremendous amount of training data, uncertainty in generalization on domain-specific tasks, and miniscule improvement with an increase in complexity of models seem to raise a concern on the features learned by the model. The utilization of relevant knowledge will aid in supervising the learning of features and facilitate explainability. The next opportunity could be to complement the implicit knowledge by KGs that already provide explicit representation with entities along with their synonyms and variants and a variety of typed relationships. Many challenges remain, such as how to represent the knowledge propagation between nodes as complex real-world relationships in a graph. Pioneers in AI are hence manipulating the structured KGs for DL with relational inductive biases (zd.net/2Jblg2A), transfer learning (interdomain knowledge sharing) and other new methods of infusing KG into ML.

Considering the challenging task of NLU that requires deciphering the unique language, semantic, and contextual characteristics, incorporating domain-specific knowledge resources, a context-aware and knowledge-enhanced computational approach will break down the content into

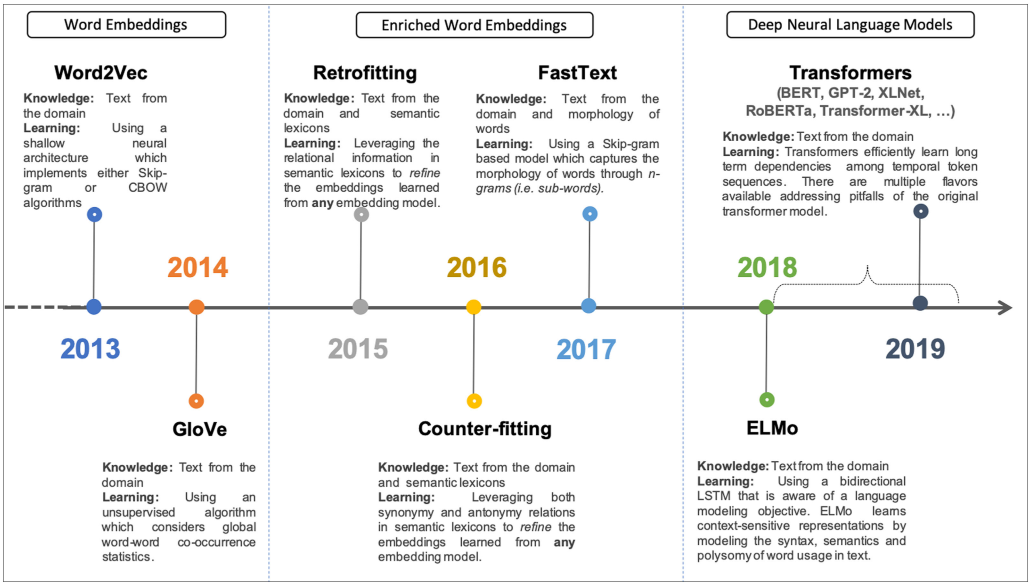

Word Embeddings
Enriched Word Embeddings
Deep Neural Language Models
Word2Vec
Knowledge: Text from the domain
Learning: Using a shallow neural architecture which implements either Skip-gram or CBOW algorithms
Retrofitting
Knowledge: Text from the domain and semantic lexicons
Learning: Leveraging the relational information in semantic lexicons to refine the embeddings learned from any embedding model.
FastText
Knowledge: Text from the domain and morphology of words
Learning: Using a Skip-gram based model which captures the morphology of words through n-grams (i.e. sub-words).
Transformers
(BERT, GPT-2, XLNet, RoBERTa, Transformer-XL, ...)
Knowledge: Text from the domain
Learning: Transformers efficiently learn long term dependencies among temporal token sequences. There are multiple flavors available addressing pitfalls of the original transformer model.
2014
2016
2018
2013
2015
2017
2019
GloVe
Knowledge: Text from the domain
Learning: Using an unsupervised algorithm which considers global word-word co-occurrence statistics.
Counter-fitting
Knowledge: Text from the domain and semantic lexicons
Learning: Leveraging both synonymy and antonymy relations in semantic lexicons to refine the embeddings learned from any embedding model.
ELMo
Knowledge: Text from the domain
Learning: Using a bidirectional LSTM that is aware of a language modeling objective. ELMo learns context-sensitive representations by modeling the syntax, semantics and polysomy of word usage in text.

Figure 1. Chronological arrangement of the existing work from the NLP domain into three paradigms by considering the degree of information captured by each model. (a) Word embeddings. (b) Enriched word embeddings using additional information. (c) Deep neural language models. Given the rapid progress in this area, we likely have not included all possible examples for 2019.

contextual building blocks that acknowledge inherent ambiguity and sparsity. To show the efficiency of such an approach, we utilized social media data (e.g., Reddit) on mental health to classify users to one of DSM-5 categories. The system showed the capability of matching the patients to mental health professionals. Our approach utilizes a zero-shot learning approach and publicly available medical knowledge graph to learn a weight matrix for modulating word vectors. Evaluations show that this approach reduces the false alarm in the classification of mental health disorders by 91% (http://bit.ly/2qU8MY1). Ananthram *et al.*[1] have utilized transportation-related ontologies to annotate events on traffic, public safety, and weather streaming as observation from citizens. The approach showed the benefits of ontologies in improving the learning performance of probabilistic graphical models.

As the infusion of knowledge in ML/DL algorithms can be at different levels of depth, we provide an overall taxonomy for knowledge infusion categorized as shallow, semi-deep, and deep infusion. We discuss each of these categories in the subsequent sections with examples.

SHALLOW INFUSION OF KNOWLEDGE

We define the first category of knowledge infusion, i.e., shallow infusion as any attempt that either completely disregards the structured knowledge or transforms them into flattened intermediate forms when used with the DL models. The two popular choices in capturing background information are first, training a shallow neural architecture or a statistical model on a large corpus and feeding the learned statistical signature as an *input* to a task-specific model[2] or second, making the task-specific model objective *directly aware* of any such background information.[15] Specifically, shallow infusion does not require the learning model to be significantly changed to ingest the external information. Rather, the external knowledge is introduced as a pretrained model or weight vectors that can be directly fed or coupled with existing neural architectures. Hence, we point out that in shallow infusion, both the information fed to a model and the method of feeding information are shallow. We highlight three alternatives from the NLP domain, as shown in Figure 1 followed by the discussions.

Word embeddings: This is the simplest form of shallow infusion. Here, the objective is to provide the model with "background" that the training data alone could not provide. The background information is available as large text corpora (for example GloVe is trained on 6B tokens) and a shallow neural network or a statistical model is trained in an unsupervised setting to capture the domain-specific meanings of words. The popular examples include but are not restricted to Word2Vec (skip-gram and CBOW algorithm) and GloVe. The representation of words as n-dimensional vectors (e.g., $n = 300$) makes them easily transferable and task-agnostic within a particular domain. As a result, numerous pretrained word embeddings are available for many languages (http://bit.do/multi-lang) and domains (http://bit.do/bionlp).

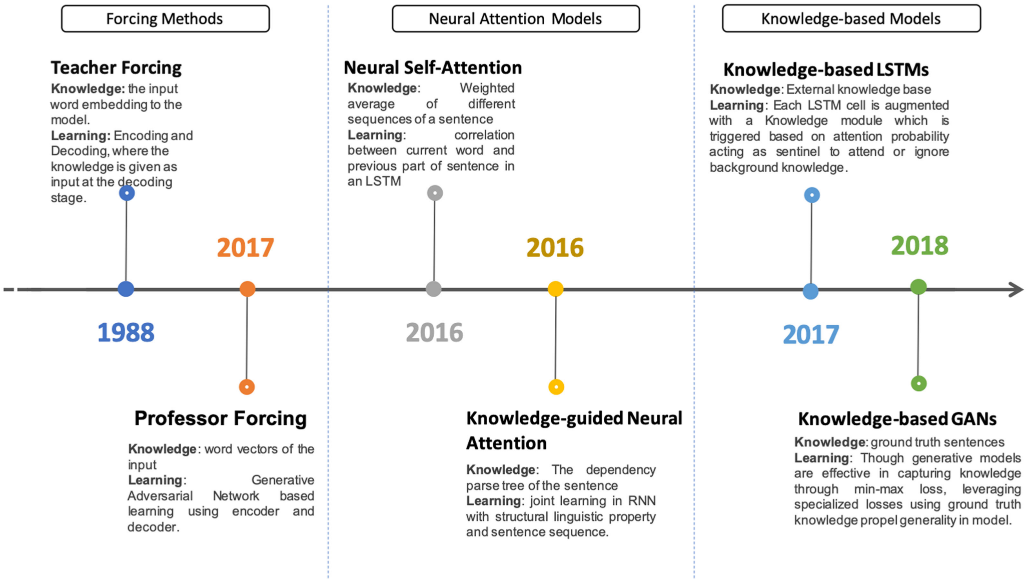
"In shallow infusion, both the external information and method of knowledge infusion is shallow."

Enriched word embeddings: In this class of algorithms, the pretrained word embeddings are *enriched* using additional information, such as domain-specific lexicons/taxonomies and morphology of words. As a postprocessing technique, "retrofitting" leverages semantic lexicons, such as WordNet in modifying the embeddings. For example, retrofitting enforces the embedding of the word "incorrect" to be in a similar vicinity to other related words, such as "wrong," "flawed," and "false" in the embedding space. "Counter-fitting," an approach similar to retrofitting, introduces synonymy and antonymy constraints to the word-relatedness when refining word embeddings. As a result, it prevents the word "inexpensive" to be closer to words, such as "pricey" and "costly" even though they are related via an antonymy relation. FastText leverages information *within* the text to improve the learned embeddings. It considers morphology of words—particularly, subword information—and represents a word as a bag of character n-grams in learning the embeddings. This allows misspelled words, rare words, and abbreviations to have a similar meaning to their original forms. Moreover, this further enables deriving embeddings for words that did not appear in the training data.

Deep neural language models: The primary difference in this class of models is the use of deep neural architectures with language modeling objective, i.e., learning to predict the next word conditioned on the given context by probabilistically modeling words in a language. ELMo marks a significant step in this direction by capturing the "context" in which a word is used in a sentence. By training a task-specific Bi-LSTM network to model the language from both forward and backward directions, ELMo represents a particular word as a combination of corresponding hidden layers. The current state-of-the-art neural language modeling is inspired by the advent of transformers—a simple, solely attention-based mechanism that disregards the need of using recurrent and convolutional neural networks. The transformer-based BERT, a model that broke records for several NLP tasks, learns to capture long-term dependencies and context by training on large amounts of text. It further fine-tunes the knowledge gained, by specifically training on a supervised-learning task. Last year has seen ground-breaking works with several transformer-based successors of BERT (e.g., RoBERTa, XLNet, and Transformer-XL) coming into light navigating the modern NLP to new directions.

SEMI-DEEP INFUSION OF KNOWLEDGE

We define the second category of knowledge infusion, i.e., *Semideep Infusion* as a paradigm that gauges the learning of a deep net and resolves the impedance mismatch by adding structural (e.g., dependency relations between words in a sentence) or symbolic (attention probability or constraints satisfaction) knowledge. Such an approach has been effective in a task-specific problem where the model is unable to learn complex representative features from the text. Further, we noticed the amalgamation of two deep learning networks is another alternative to bring together structural and sequential learning for improving the prediction (Yin *et al.*). We categorize different perspectives of semideep infusion of knowledge in the deep neural networks outlined for various NLP/NLU tasks (e.g. event detection, user classification, relationship extraction, reading comprehension, etc.).



Teacher Forcing
Knowledge: the input word embedding to the model.
Learning: Encoding and Decoding, where the knowledge is given as input at the decoding stage.

Professor Forcing
Knowledge: word vectors of the input
Learning: Generative Adversarial Network based learning using encoder and decoder.

Neural Self-Attention
Knowledge: Weighted average of different sequences of a sentence
Learning: correlation between current word and previous part of sentence in an LSTM

Knowledge-guided Neural Attention
Knowledge: The dependency parse tree of the sentence
Learning: joint learning in RNN with structural linguistic property and sentence sequence.

Knowledge-based LSTMs
Knowledge: External knowledge base
Learning: Each LSTM cell is augmented with a Knowledge module which is triggered based on attention probability acting as sentinel to attend or ignore background knowledge.

Knowledge-based GANs
Knowledge: ground truth sentences
Learning: Though generative models are effective in capturing knowledge through min-max loss, leveraging specialized losses using ground truth knowledge propel generality in model.

Figure 2. Ordering of existing work that relates to our definition of Semideep Infusion of Knowledge. We categorize the process of semideep infusion into three paradigms: (a) forcing methods, (b) neural attention models, and (c) knowledge-based models. Given the rapid progress in this area, we likely have not included all possible examples for 2019.

Teacher/professor forcing: In a deep learning framework comprising of an autoencoder, the capability of a decoder is enhanced through teacher forcing. In this procedure, the target labels (nonbinary rather structured sentences) are fed word by word while training the decoder part of the autoencoder. The vectorized representation of the input on which decoder tries to learn is provided by the encoder. The procedure was first discussed by Williams *et al.* and has shown improvement in machine translation, entity extraction, and negation detection tasks (Lamb *et al.*). Understanding the procedure of teacher forcing, we identified two critical issues: first, the representation provided by the encoder is not gauged in the teacher forcing method, and second, the model memorizes the patterns of the input and is difficult to perform transfer learning with the trained model. For example, consider learning of an autoencoder over "harassment dataset from social media" through teacher forcing, it is uncertain for the model to perform well on a near-related problem of "radicalization of social media." It is because of poor contextualization and adaptability of the model. Kursuncu *et al.* leveraged domain-specific perspective models in enriching the representation of extremist's communication on social media.[8] The approach provided the necessary knowledge required by a model to minimize the false alarm. In the context of the problem of "harassment on social media," a potential improvement in a machine learning model has been made through the infusion of cyberbullying vocabulary knowledge.

A teacher forced model is able to learn the correct representation of the input through the below methods:

- Redundancy: In this learning process, the model is monitored for the information loss through backpropagation and is replenished through replicating the input to the layers. Methods, such as skip connections or highway connections follow such a methodology.
- Curriculum learning: A variation of forced learning is to introduce outputs generated from prior time steps during training to encourage the model to learn how to correct its own mistakes.

In the teacher forcing paradigm, during inference, the conditioning context may diverge during training when ground truth labels are given as input. Since the encoder acts as a generator and decoder behave like discriminator, their independent functioning affects the model performance. Further, the incorporation of the knowledge is on the decoder side independent of the encoder. Hence, it is challenging to quantify the loss of information incurred on the encoder side. Our proposed approach on *Deep Infusion* regulates first, *where in model, the latent weights are wrongly enforced* and second, *how to adjust the weights leveraging external human-curated graphical knowledge source*.

Neural attention models (NAM): Attention models highlights particular features that are important for the pattern recognition/classification based on a hierarchical architecture of the content. The manipulation of attentional focus is effective in solving the real-world problems involving massive amounts of data (Sun *et al.* 2017). On the other hand, some applications demonstrate the limitation of attentional manipulation in a set of problems such as sentiment (mis)classification and suicide risk,[5] where feature presence is inherently ambiguous, just as in the radicalization problem. For example, in the suicide risk prediction task, references to the suicide-related terminology appear in the social media posts of both victims as well as supportive listeners, and the existing NAMs fail to capture semantic relations between terms to help differentiate the suicidal user from a supportive user. To overcome such limitations in a sentiment classification task, Vo *et al.*[13] have augmented sentiment scores in the feature set for enhancing the learned representation and modified the loss function to respond to the values of the sentiment score during learning. However, Sheth *et al.*[11] have pointed out the importance of using domain-specific knowledge especially in cases where the problem is complex. In an empirical study, Bian *et al.*[3] showed the effectiveness of combining richer semantics from domain knowledge with morphological and syntactic knowledge in the text, by modeling knowledge assistance as an auxiliary task that regularizes learning of the main objective in a deep neural network.

> "In semideep infusion, external knowledge is involved through attention mechanism or learnable knowledge constraints acting as a sentinel to guide model learning."

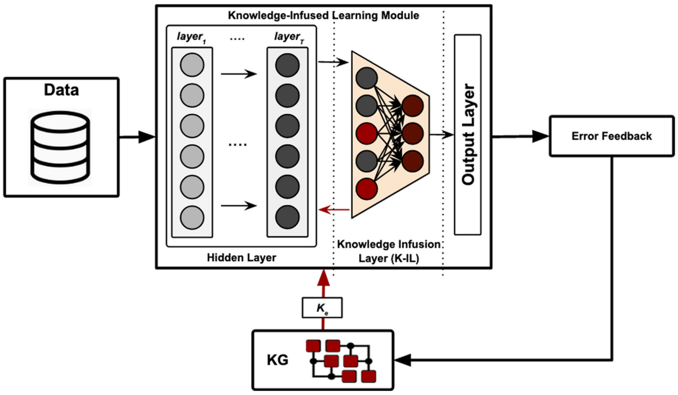


Figure 3. Representations of data are generated, and domain knowledge amplifies the significance of specific important concepts that are otherwise missed in the learning model. Classification error and KG determine the need for infusing knowledge. The Knowledge Infusion Layer incorporates the knowledge in the latent representation before output layer.

Learnable knowledge constraints: Professor forcing forms an architecture where the encoder (generator) competes with the decoder (discriminator) in improving the outcome, thus forming an adversarial network. Further, the improvement in the learning occurs by acting as a posterior regularizer and allowing the possibility of including rich structured domain knowledge. However, in professor forcing, if knowledge constraints need to be infused, they need to be done *a priori* and not iteratively while learning. A recent study from Hu *et al.*[6] focuses on infusing the knowledge as constraints in such an adversarial network by optimizing the Kullback–Leibler (KL) divergence. However, the knowledge gathered for infusion is part of the dataset and does not exploit a human-curated knowledge graph. Further, the study relates to our objective by monitoring the KL divergence. However, it does not show an appropriate methodology on adding the relevant knowledge, which is quantified from the KL score. However, in our *deep infusion* paradigm (see Figure 3), we aim at defining the quantification and inclusion of relevant knowledge to deep models to minimize the learning time and false alarm rate.

Graph neural network (GNN): Graph neural network is a type of neural network that directly operates on the graph structure (Scarselli *et al.* 2008). A typical application of GNN is a node classification. Essentially, every node in the graph is associated with a label, and we want to predict the label of the nodes without ground-truth. In this process, the model generates importance score for each node and the connection weights form the weights of the relationship between the nodes. In this and a similar study,[14] the GNN framework can be seen as leveraging the structural property of the KG and quantifying itself using the input data. However, the framework is restricted to the labels in the input dataset and their interrelationships. Further, the GNN does not exploit the structural property and taxonomic relationships of the KG in identifying the relevant knowledge that can be applied to the learning of the neural network. Further, the hidden nodes in GNN are unaware abstractions corresponding to a stratified knowledge in a KG, thus the relationships between the labels are not well contextualized.

Tree LSTMs: LSTMs are sequential models, whereas the sentences in the input corpus follow a grammatical tree structure (dependency or constituency). Hence, it is important to learn the contextual representation of the input following the same tree structure. The Tree LSTMs (Tai *et al.* 2015) replaces the nodes in the graph with LSTMs cells and vector representation of the words/phrases is given as input. This model considers structural (syntactic) property of the input, but the domain knowledge is ignored.

A recent study from Yang *et al.* 2017 utilizes external knowledge bases (e.g., WordNet, NELL) to improve the performance of BiLSTMs by minimizing task-specific feature engineering. Particularly, the study focused on improving entity and event extraction. Knowledge-based LSTM proposed in the study comprises of an attention mechanism that acts as a sentinel to guide the model in deciding whether to use external knowledge and adaptively decide the level of abstractness in the information. Although the proposed architecture uses external knowledge base as a separate component for each LSTM cell, it is uncertain *how much of the external knowledge needs to be incorporated* and to *what level of abstraction the traversing of the knowledge base needs to be done* to fulfill the information loss in the learning process.

Papers on DL techniques cited:

- (Word2Vec): T. Mikolov *et al.,* “Distributed representations of words and phrases and their compositionality,” *Proc. 26th Int. Conf. Neural Inf. Process. Syst.,* 2013, pp. 3111–3119.
- (GloVe): J. Pennington *et al*., “Glove: Global vectors for word representation,” *in Proc. Conf. Empirical Methods Natural Lang. Process.,* 2014, pp. 1532–1543.
- (Retrofitting): M. Faruqui *et al.,* “Retrofitting word vectors to semantic lexicons,” *in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.,* 2015, pp. 1606–1615.
- (Counter-fitting): N. Mrkšić *et al.,* “Counter-fitting word vectors to linguistic constraints,” *in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.,* 2016, pp.142–148.
- (FastText): P. Bojanowski *et al.,* “Enriching word vectors with subword information,” in Proc. *Trans. Assoc. Comput. Linguistics*, 2017, pp. 135–146.
- (ELMo): M. E. Peters *et al*., “Deep contextualized word representations,” *in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.,* 2018.
- (Transformers): A. Vaswani *et al*., “Attention is all you need,” *Proc. 31st Int. Conf. Neural Inf. Process. Syst.,* 2017, pp. 6000–6010.
- (BERT): J. Devlin *et al.,* “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*.
- (GPT-2): A. Radford *et al.,* “Language models are unsupervised multitask learners,” OpenAI Blog, 2019.
- (XLNet): Z. Yang *et al*., “XLNet: Generalized autoregressive pre-training for language understanding,” 2019, *arXiv:1906.08237*.
- (RoBERTa): Y. Liu *et al*., “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, *arXiv:1907.11692*.
- (Teacher Forcing): R. J. Williams *et al*., “A learning algorithm for continually running fully recurrent neural networks,” *Neural Comput.,* vol. 1, no. 2, Jun. 1989.
- (Professor Forcing): A. Lamb *et al*., “Professor forcing: A new algorithm for training recurrent networks,” *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4608–4616.
- (NAM): M. R. Alexander *et al*., “A neural attention model for abstractive sentence summarization,” in *Proc. Conf. Empirical Methods Natural Lang. Process.,* 2015, pp. 379–389.
- (Tree-LSTM): K. S. Tai *et al.*, “Improved semantic representations from tree-structured long short-term memory networks,” *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 1556–1566.
- (GNN): F. Scarselli *et al*., “The graph neural network model,” *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2008.
- (NAM): K. Yi *et al.* “Knowledge-based recurrent attentive neural network for small object detection,” 2018, *arXiv:1803.05263*.
- (TransE): Z. Wang *et al*., “Knowledge graph embedding by translating on hyperplanes,” in *Proc. 28th AAAI Conf. Artif. Intell.,* 2014, pp. 1112–1119.

- (KBLSTM): B. Yang *et al.*, “Leveraging knowledge bases in LSTMs for improving machine reading,” *in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1436–1446.

DEEP INFUSION OF KNOWLEDGE

We define the third category of knowledge infusion, i.e., *Deep Infusion of Knowledge* as a paradigm that couples the latent representation learned by deep neural networks with the KGs exploiting the semantic relationships between entities. We aim to: first, quantify the information loss, second, identify the relevant knowledge at an appropriate level of abstraction, third, appropriately combine representation of identified concepts in KGs with a latent representation of data. The existing research shows the contribution of incorporating external knowledge in machine learning, this incorporation mostly takes place before or after the actual learning process. We argue that deep infusion within the latent layers of neural networks will boost the performance of neural networks as an integral component of the AI models deployed in applications. With a deep infusion of such structured knowledge, it will reveal patterns that are missed by shallow and semideep infusion because of sparse feature occurrence, ambiguity, and noise. This approach will allow to accomplish the infusion of declarative domain knowledge in the latent layers of neural networks.

> “It would be useful to use a stratified representation of knowledge representing different levels of abstractions. As we understand the level of abstraction represented by different layers in a deep learning model, we can look to transfer knowledge that aligns with the corresponding layer in the layered learning process.”

Among current state-of-the-art works, Yi *et al.* 2018 have introduced a knowledge-based recurrent attention neural network (KB-RANN) that modifies the attentional mechanism by incorporating domain knowledge to make the model generalize better. However, their domain-knowledge is statistically derivable from the existing data, without capturing exceptions, anomalies, and irregularities, which are sparse but important knowledge that helps characterizing semantic cues and nuances. The studies for incorporating knowledge in a deep learning process have not involved structured knowledge in the form of KGs. On the other hand, Casteleiro *et al.* 2018 recently showed how the cardiovascular disease ontology provided context and reduced ambiguity, improving performance on a synonym detection task. Researchers employed embeddings of entities in a KG, derived through Bi-LSTMs, to enhance the efficacy of neural attention models. Looking ahead, given that KGs use a rich graphical representation, we believe that graphical neural networks will provide richer ways to align knowledge with the learning process and support infusion while maintaining the richness of knowledge representation, such as link (relationship) semantics. These existing studies utilized external knowledge *after* the representation has been generated by neural language models, *rather than within the deep neural network*. We argue that a learning framework that incorporates domain knowledge within the latent layers of neural networks for modeling will improve performance in a holistic manner.

In healthcare, for example, infusing knowledge would mean incorporating rich domain knowledge captured in manually curated medical KGs (e.g., UMLS, ICD-10 and DataMed) while not losing all the abstractions and context (e.g., a term used in “family history” has a different meaning than the same term used in “impression and plan” in an EMR), taxonomic and named relationships, and complex and compound entities (e.g., “adenomatous hyperplasia of endometrium” is a single entity, and any system that thinks this related to hyperplasia or endometrial would be using incorrect semantics). In DL for the NLP, knowledge corresponding to linguistic aspects or components (words, entities and relationships, modifiers, phrases recognized by parsed trees, etc.) will be incorporated at different layers in the learning process. In a task like deep learning used for image processing, the knowledge for texture is best incorporated at an intermediate layer that corresponds to the abstraction of texture and best utilize it. As each layer in a neural network architecture produces a latent representation that is transmitted between hidden layers, the infusion of knowledge during this learning process raises the relevant research questions: first, how to decide whether to infuse knowledge or not at a particular stage in learning between layers, and how to measure the incorporation of knowledge? Second, how to merge latent representations with knowledge representations, and how to propagate the knowledge through the learned representation? While these research questions require further investigations, we believe that developing functions in a neural network architecture with respect to representations of external knowledge will be critical. As the goal is to infuse knowledge within the neural network, the architecture can be designed as follows: first, before the output layer (see Figure 3) and second, between the hidden layers.

While it is essential to have an appropriate design for neural network architecture, the creation of appropriate knowledge representation to be infused in to the neural networks is also crucial. As a representation of knowledge in the KG can be typically generated as embedding vectors, it still does not truly represent the semantics and requires further investigation to reflect the power of knowledge in a KG with its relationships.[8] Specific contextual models and/or more generic models can be utilized to create an embedding of each concept and their relations in a KG through the proximity using appropriate distance measures (e.g., Least Common Subsumer). Further, existing knowledge embedding models can be utilized such as TRANS-E, TRANS-H, and HOLE for the creation of embeddings from KGs.

As we argue that knowledge infusion can occur between hidden layers or just before the output layer. Kursuncu *et al.*[7,8] detailed an initial approach for the Knowledge Infusion Layer for the scenario, which takes place just before the output layer. In neural language models, the output layer (e.g., SoftMax) estimates the error to be back-propagated. Each epoch generates an error, which is incrementally reduced, and it is back-propagated until the model reaches a saddle point in the local minima. The error represents the difference between the actual and predicted labels. Two specific functions were introduced as an initial approach to optimize the loss function with respect to the KL divergence and merge the latent vectors from the hidden layers and the knowledge embedding. This approach estimates the divergence between the latent representations and knowledge representation, to determine the differential knowledge to be infused. Further, modulation of the knowledge-infused learned weight matrix and latent representation will be critical and will need further investigations.

DISCUSSION

In this article, we overviewed the continuing progress toward using and incorporating structured knowledge to develop increasingly more powerful learning techniques. Future advances in this area will integrate top-down and bottom-up processing, moving AI techniques closer to how cognitive scientists believe human brain's function.

ACKNOWLEDGMENTS

The authors would like to thank S. Padhee, T. K. Prasad, and B. Srivastava for their reviews and comments.

REFERENCES

1. P. Anantharam *et al.*, "Extracting city traffic events from social streams," *ACM Trans. Intell. Syst. Technol.*, vol. 23, no. 60, 2015, Art. no. 43.
2. M. Baroni *et al.*, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proc 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 238–247.
3. J. Bian *et al.*, "Knowledge-powered deep learning for word embedding," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2014, pp. 132–148.
4. P. M. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.
5. M. Gaur *et al.*, "Knowledge-aware assessment of severity of suicide risk for early intervention," in *Proc. World Wide Web Conf.*, 2019, pp. 514–525.
6. Z. Hu *et al.*, "Deep generative models with learnable knowledge constraints," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 10522–10533.

7. U. Kursuncu *et al.*, "Knowledge infused learning (K-IL): Towards deep incorporation of knowledge in deep learning," in *Proc. AAAI Spring Symp., Combining Mach. Learn. Know. Eng.*, Palo Alto, CA, USA, 2020, arXiv:1912.00512.
8. U. Kursuncu *et al.*, "Modeling islamist extremist communications on social media using contextual dimensions: religion, ideology, and hate," *Proc ACM Human-Comput. Interact.*, vol. 3, 2019, Art. no. 151.
9. M. A. Casteleiro *et al.*, "Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature," *J. Biomed. Semantics*, 2018.
10. A. Sheth *et al.*, "Semantics for the semantic web: The implicit, the formal and the powerful," *Int. J. Semantic Web Inf. Syst.*, vol. 1, 2005, Art. no. 18.
11. A. Sheth *et al.*, "Knowledge will propel machine understanding of content: Extrapolating from current examples," in *Proc Int. Conf. Web Intell.*, 2017, pp. 1–9.
12. C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.
13. K. Vo *et al.*, "Combination of domain knowledge and deep learning for sentiment analysis," in *Proc. Int. Workshop Multi-Disciplinary Trends Artif. Intell.*, 2017, *arXiv:1806.08760.*
14. X. Wang *et al.*, "Explainable reasoning over knowledge graphs for recommendation," in *Proc AAAI Conf. Artif. Intell.*, 2019.
15. J. Xu *et al.*, "A semantic loss function for deep learning with symbolic knowledge," in *Proc. Int. Conf. Mach. Learn.*, 2018.

Amit Sheth is the director of the Artificial Intelligence Institute, University of South Carolina, Columbia, SC, USA. Contact him at: amit@knoesis.org.

Ugur Kursuncu is a postdoctoral fellow. Contact him at: kursuncu@mailbox.sc.edu.

Manas Gaur is working toward the Ph.D. degree with Prof. Sheth. Contact him at: mgaur@email.sc.edu.

Ruwan Wickramarachchi is working toward the Ph.D. degree with Prof. Sheth. Contact him at: ruwan@email.sc.edu.