

Knowledge Infused Learning (K-IL): Towards Deep Incorporation of Knowledge in Deep Learning

Ugur Kursuncu*, Manas Gaur*, Amit Sheth

AI Institute, University of South Carolina
Columbia, SC, USA

{kursuncu@mailbox.sc.edu, mgaur@email.sc.edu, amit@sc.edu}

Abstract

Learning the underlying patterns in data goes beyond instance-based generalization to external knowledge represented in structured graphs or networks. Deep learning that primarily constitutes neural computing stream in AI has shown significant advances in probabilistically learning latent patterns using a multi-layered network of computational nodes (i.e., neurons/hidden units). Structured knowledge that underlies symbolic computing approaches and often supports reasoning, has also seen significant growth in recent years, in the form of broad-based (e.g., DBPedia, Yago) and domain, industry or application specific knowledge graphs. A common substrate with careful integration of the two will raise opportunities to develop neuro-symbolic learning approaches for AI, where conceptual and probabilistic representations are combined. As the incorporation of external knowledge will aid in supervising the learning of features for the model, deep infusion of representational knowledge from knowledge graphs within hidden layers will further enhance the learning process. Although much work remains, we believe that knowledge graphs will play an increasing role in developing hybrid neuro-symbolic intelligent systems (bottom-up deep learning with top-down symbolic computing) as well as in building explainable AI systems for which knowledge graphs will provide scaffolding for punctuating neural computing. In this position paper, we describe our motivation for such a neuro-symbolic approach and framework that combines knowledge graph and neural networks.

Introduction

Data-driven bottom-up machine/deep learning (ML) and top-down knowledge-driven approaches to creating reliable models, have shown remarkable success in specific areas, such as search, speech recognition, language translation, computer vision, and autonomous vehicles. On the other hand, they have had limited success in understanding and deciphering contextual information, such as detection of abstract concepts in online/offline human interactions. Current

challenges in the translation of research methods and resources into practice often draw from a class of rarely studied problems that do not yield to contemporary bottom-up ML methods. Policymakers and practitioners assert serious usability concerns that constrain adoption, notably in high-consequence domains (Topol 2019). In most cases, data-dependent ML algorithms require high computing power and large datasets, where the crucial signals may still be sparse or ambiguous, threatening precision (Cheng 2018). Moreover, the ML models that are deployed in the absence of transparency and accountability (Rudin 2019) and trained on biased datasets, can lead to grave consequences, such as potential social discrimination and unfair treatment (Olteanu et al. 2019). Further, the potentially severe implications of false alarms in an ML-integrated real-world application may affect millions of people (Kursuncu et al. 2019a; Kursuncu 2018).

The fundamental challenges are common to a majority of problems in a variety of domains with real world impact. Specifically, these challenges are: (1) *dependency on large datasets* required for bottom-up, data-dependent ML algorithms (Valiant 2000; De Palma, Kiani, and Lloyd 2019), (2) *bias in the dataset*, enabling the model to potentially cause social discrimination and unfair treatment, (3) *multidimensionality, ambiguity and sparsity*, as the data involves unconstrained concepts and relationships with meaning from different contextual dimensions of the content such as religion, history and politics (Kursuncu et al. 2019a; Kursuncu 2018). Further, the limited number of labeled instances available for training may fail to represent the true nature of concepts and relationships in data sets, leading to ambiguous or sparse true signals (4) the lack of information traceability for model *explainability*, (5) *the coverage of information* specific to a domain that would be missed otherwise, (6) *the complexity of model architecture in time and space*¹, and (7) *false alarms* in model performance. Consequently, we believe standard separate knowledge graph KG and ML methods are vulnerable to deduce or learn spurious concepts and relationships that appear deceptively good on a KG or training datasets, yet do not provide adequate re-

*Equally Contributed

Copyright © 2020 held by the author(s). In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice (AAAI-MAKE 2020). Stanford University, Palo Alto, California, USA, March 23-25, 2020. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.theguardian.com/commentisfree/2019/nov/16/can-planet-afford-exorbitant-power-demands-of-machine-learning>

sults when the data set contains contextual and dynamically changing concepts and relations.

In this position paper, we describe innovations that will operationalize more abstract models built upon the characteristics of a domain to render them computationally accessible within neural network architectures. We propose a neuro-symbolic method, *knowledge-infused learning* that measures information loss in latent features learned by neural networks through KGs with conceptual and structural relationship information, for addressing the aforementioned challenges. The infusion of knowledge during the representation learning phase raises the following central research questions: (i) How do we decide whether to infuse knowledge or not, at a particular stage while learning between layers, and how to quantify knowledge to be infused? (ii) How to merge latent representations between layers with external knowledge representations, and (iii) How to propagate the knowledge through the learned latent representation? Considering the future deployment of AI in applications, the potential impact of this approach is significant. As stated in (Karpathy 2015), the deeper the network, the denser the representation and better the learning. A large number of parameters and the layered nature of neural networks make them modifiable based on specific problem characteristics. However, the challenges (1, 3, 5 and 7) make neural networks vulnerable to the sudden appearance of relevant-but-sparse or ambiguous features, in often noisy big data (Valiant 2000; De Palma, Kiani, and Lloyd 2019; Kursuncu et al. 2019b). On the other hand, KG-based approaches structure search within a feature space defined by domain experts. To compensate for the vulnerability of the aforementioned challenges, incorporating knowledge to the learned representation in principled fashion is required. A promising approach is to base this on a measurable discrepancy between the knowledge captured in the neural network and external resources.

Computational modeling coupled with knowledge infusion in a neural network will disambiguate important concepts defined in a KG with their different semantic meanings through its structural relations. Knowledge infusion will redefine the emphasis of sparse but essential, and irrelevant but frequently occurring terms and concepts, boosting recall without reducing precision. Further, it will provide explanatory insight into the model, robustness to noise and reduce dependency on frequency in the learning process. This neuro-symbolic learning approach will potentially transform existing methods for data analysis and building computational models. While the impact of this approach is transferable (and replicable) to a majority of domains, the explicit implications are particularly apparent for social science (Kursuncu et al. 2019a) and healthcare domains (Gaur et al. 2018).

Related Work

As the incorporation of knowledge has been explored in various forms in prior research, in this section, we describe the methodologies and applications specifically related to knowledge-infused learning: Neural language models, neural attention models, knowledge based neural networks, all

of which utilize external knowledge before/after the representation has been generated.

Neural Language Models (NLMs)

NLMs are a category of neural networks capable of learning sequential dependencies in a sentence, and preserve such information while learning a representation. In particular, LSTM (Long Short Term Memory) networks (Hochreiter and Schmidhuber 1997) have emerged from the failure of RNNs (Recurrent Neural Networks) in remembering long-term information. Concerning the loss of contextual information while learning, (Cho et al. 2014) proposed a context-feed forward LSTM architecture in which context is learned by the previous layer merged with forgetting and modulation gates of the next layer. However, if erroneous contextual information is learned in previous layers, it is difficult to correct (Masse, Grant, and Freedman 2018), which is a problem magnified by noisy data and content sparsity (e.g. Twitter, Reddit, Blogs).

As the inclusion of structured knowledge (e.g., Knowledge Graphs) in deep learning, improves information retrieval (Sheth and Kapanipathi 2016), prior research has shown the significance of knowledge in the pursuit of improving NLMs, such as in commonsense reasoning (Liu and Singh 2004). The transformer NLMs such as BERT, (Devlin et al. 2018) (including its variants BioBERT and SciBERT), are still data dependent. BERT has been utilized in hybrid frameworks such as (Scarlina, Pasini, and Navigli 2020) in the creation of sense embeddings using BabelNet and NASARI. (Liu et al. 2019a) proposed K-BERT, that enriches the representations by injecting the triples from KGs into the sentence. As this incorporation of knowledge for BERT takes place in the form of attention, we consider the K-BERT as semi-deep infusion (Sheth et al. 2019). Similarly, ERNIE (Sun et al. 2019) incorporated external knowledge to capture lexical, syntactic, and semantic information, enriching BERT.

Neural Attention Models (NAM)

NAM (Rush, Chopra, and Weston 2015) highlights particular features that are important for pattern recognition/classification based on a hierarchical architecture. The manipulation of attentional focus is effective in solving real world problems involving massive amounts of data (Halevy, Norvig, and Pereira 2009; 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 (Maurya 2018) and suicide risk (Corbitt-Hall et al. 2016), where feature presence is inherently ambiguous, just as in the online radicalization problem (Kursuncu et al. 2019a). For example, in the suicide risk prediction task, references to suicide-related terminology appear in social media posts of both victims as well as supportive listeners, and the existing NAMs fail to capture semantic relations between terms that help differentiate the suicidal user from a supportive user (Gaur et al. 2019). To overcome such limitations in a sentiment classification task, (Vo et al. 2017) adds sentiment scores into the feature set for enhancing the learned representation and

modifies the loss function to respond to values of the sentiment score during learning. However, (Sheth et al. 2017; Kho et al. 2019) have pointed out the importance of using domain-specific knowledge especially in cases where the problem is complex in nature (Perera et al. 2016). (Bian, Gao, and Liu 2014) has empirically demonstrated the effectiveness of combining richer semantics from domain knowledge with morphological and syntactic knowledge in the text, by modeling knowledge as an auxiliary task that regularizes the learning of the main objective in a deep neural network.

Knowledge-based Neural Networks

(Yi et al. 2018) introduced a knowledge-based, recurrent attention neural network (KB-RANN) that modifies the attentional mechanism by incorporating domain knowledge to improve model generalization. However, their domain-knowledge is statistically derivable from the input data itself and is analogous to merely learning an interpolation function over the existing data. (Dugas et al. 2009) proposed a modification in the neural network by adopting Lipschitz functions for its activation function. (Hu et al. 2016) proposed a combination of deep neural networks with logic rules by employing knowledge distillation procedure (Hinton, Vinyals, and Dean 2015) of transferring the learned tacit knowledge from larger neural network, to the weights of the smaller neural network in data-limited settings. These studies for incorporating knowledge in a deep learning framework have not involved declarative knowledge structures in the form of KGs (e.g., DBpedia) (Chen et al. 2019). However, (Casteleiro et al. 2018) recently showed how the Cardiovascular Disease Ontology (CDO) provided context and reduced ambiguity, improving performance on a synonym detection task. (Shen et al. 2018) employed embeddings of entities in a KG, derived through Bi-LSTMs, to enhance the efficacy of NAMs. (Sarker et al. 2017) presented a conceptual framework for explaining artificial neural networks' classification behavior using background knowledge on the semantic web. (Makni and Hendler) explained a deep learning approach to learn RDFS² rules from both synthetic and real-world semantic web data. They also claim their approach improves the noise-tolerance capabilities of RDFS reasoning.

All of the frameworks in the above subsections utilized external knowledge before or after the representation has been generated by NAMs, *rather than within the deep neural network* as in our approach (Sheth et al. 2019). We propose a learning framework that infuses domain knowledge within the latent layers of neural networks for modeling.

Preliminaries

Symbolic representation of a domain, besides its probabilistic representation, is crucial for neuro-symbolic learning. In our approach, we propose to homogenize symbolic information from KGs (see Section Knowledge Graphs) and contextual neural representations (see Section Contextual Modeling), in neural networks.

²<https://www.w3.org/2001/sw/wiki/RDFS>

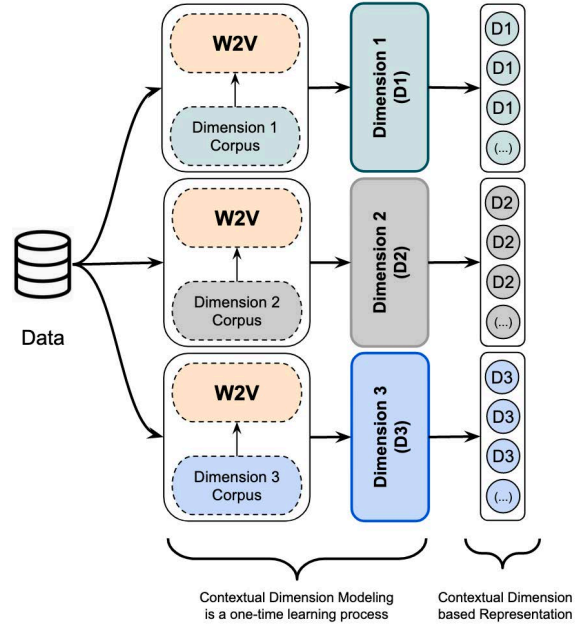


Figure 1: Contextual Dimension Modeling Diagram (Kursuncu et al. 2019a) Embedding algorithm above (W2V: Word2Vec) can be replaced by other algorithms such as BERT. For each dimension, a specific corpus is utilized to create the model and the generated representation of content is concatenated. Generating the three contextual dimension representations of a social media post will emphasize the weights of such essential lexical cues.

Knowledge Graphs

A Knowledge graph (KG) is a conceptual model of a domain that stores and structures declarative knowledge in a human and machine-readable format, constituting factual ground truth and embodying a domain ontology of objects, attributes, and relations. KGs rely on symbolic propositions, employing generic conceptual relationships in taxonomies, partonomies and specific content with labeled links. Examples include DBpedia, UMLS, and ICD-10. The factual information about the domain is represented in the form of instances (or individuals) of those concepts (or classes) and relationships (Gruber 2008; Sheth and Thirunarayan 2012). Therefore, a domain can be described or modeled through KGs in a way that both computers and humans can understand. As KGs differentiate contextual nuances of concepts in the content, they play a key role in our framework with extensive use by several functions.

Contextual Modeling

Capturing contextual cues in the language is crucial in our approach; hence, we utilize NLMs to generate embeddings of the content. Recent embedding algorithms have emerged to create such representations such as Word2Vec (Goldberg and Levy 2014), GLoVe (Pennington, Socher, and Manning 2014), FastText (Athiwaratkun, Wilson, and Anandkumar 2018) and BERT (Devlin et al. 2018).

Modeling context-sensitive problems in different domains (e.g., healthcare, cyber social threats, online extremism and harassment), depends heavily on carefully designed features to extract meaningful information, based on characteristics of the problems and a ground truth dataset. Moreover, identifying these characteristics and differentiating the content requires different levels of granularity in the organization of features. For instance, in the problem of online Islamist extremism, the information being shared in social media posts by users in extremism-related social networks displays an intent that depends on the user’s type (e.g., recruiter, follower). Hence, as these user types show different characteristics (Kursuncu et al. 2018), for reliable analysis, it is critical to consider different contextual dimensions (Kursuncu et al. 2019a; Kursuncu 2018). Moreover, the ambiguity of diagnostic terms (e.g., jihad) also mandates representation of terms in different contexts. Hence, to better reflect these differences, creating multiple models enables us to represent the multiple contextual dimensions for a reliable analysis. Figure 1 details the contextual dimension modeling workflow.

A Proposed Comprehensive Approach

Although the existing research (Gaur et al. 2018; Bhatt et al. 2018a) shows the contribution of incorporating external knowledge in ML, this incorporation mostly takes place before or after the actual learning process (e.g., feature extraction, validation); thus remaining shallow. We believe that deep knowledge infusion, within the hidden layers of neural networks, will greatly improve the performance by: (i) reducing false alarm and information loss, (ii) boosting recall without sacrificing precision, (iii) providing finer granular representation, (iv) enabling explainability (Islam et al. 2019; Kursuncu et al. 2019c) and (v) reducing bias. Specifically, we believe that it will become a critical and integral component of AI models that are integrated in deployed tools, e.g. in healthcare, where domain knowledge is crucial and indispensable in decision making processes. Fortunately, these domains are rich in terms of their respective machine-readable knowledge resources, such as manually curated medical KGs (e.g., UMLS (McInnes, Pedersen, and Pakhomov 2009), ICD-10 (Brouch 2000) and DataMed (Ohno-Machado et al. 2017)). In our prior research (Gaur et al. 2018), we utilized ML models coupled with these KGs to predict mental health disorders among 20 Mental Disorders (defined in the DSM-5) for Reddit posts. Typical approaches for such predictions employ word embeddings, such as Word2Vec, resulting in sub-optimal performance when they are used in domain-specific tasks. We have incorporated knowledge into the embeddings of Reddit posts by (i) using Zero Shot learning (Palatucci et al. 2009), (ii) *modulating* (e.g., re-weighting) their embeddings, similar to NAMs, and obtained a significant reduction in the false alarm rate, from 13% (without knowledge) to 2.5% (with knowledge). In another study, we have leveraged the domain knowledge in KGs to validate model weights that explain diverse crowd behavior in the Fantasy Premier League participants (FPL) (Bhatt et al. 2018b). However, very little previous work has tried to integrate such functional knowledge to

an existing deep learning framework.

We propose to further develop an innovative deep knowledge-infused learning approach that will reveal patterns that are missed by traditional approaches because of sparse feature occurrence, feature ambiguity and noise. This approach will support the following integrated aims: (i) Infusion of Declarative Domain Knowledge in a Deep Learning framework, and (ii) Optimal Sub-Knowledge Graph Creation and Evolution. The overall architecture in Figure 2 guides our proposed research on these two aims. Our methods will disambiguate important concepts defined in the respective KGs with their different semantic meanings through its structural relations. Knowledge infusion will re-define the emphasis of *sparse-but-essential*, and *irrelevant-but-frequently-occurring* terms and concepts, boosting recall without reducing precision.

Knowledge-Infused Learning

Each layer in a neural network architecture produces a latent representation of the input vector (h_t). The infusion of knowledge during the representation learning phase raises the following central research questions: **R1:** How do we decide whether to infuse knowledge or not, at a particular stage while learning between layers, and how to quantify knowledge to be infused? **R2:** How to merge latent representations between layers with external knowledge representations, and **R3:** How to propagate the knowledge through the learned latent representation? We propose to define two functions to address these two questions: *Knowledge-Aware Loss Function (K-LF)* and *Knowledge Modulation Function (K-MF)*, respectively.

Configurations of neural networks can be designed in various ways depending on the problem. As our aim is to infuse knowledge within the neural network, such an operation can take place (i) before the output layer (e.g., Soft-Max), (ii) between hidden layers (e.g., reinforcing the gates of an NLM layer, modulating the hidden states of NLM layers, Knowledge-driven NLM dropout and recurrent dropout between layers). To illustrate (i), we describe our initial approach to neural language models that infuses knowledge before the output layer, which we believe will shed the light towards a reliable and robust solution with more research and rigorous experimentations.

Seeded Sub-Knowledge Graph The Seeded Sub-Knowledge Graph, is a subset of KGs, which participate broadly in our technical approach. Generic KGs (e.g., DBpedia (Bizer et al. 2009), YAGO2 (Hoffart et al. 2013), Freebase (Bollacker et al. 2008)) may contain over a million entities and close to a billion relationships. Using the entire graph of linked data on the web can cause; (1) unnecessary computation and (2) noise due to irrelevant knowledge, and has sometimes failed to benefit intelligent application (Roy, Park, and Pan 2017). However, real-world problems are domain-specific and require only a relevant (sub) portion of the full graph. Creation of a Seeded Sub-KG (Lalithsena 2018) based on a ground truth dataset is needed, to represent a particular domain using information-theoretic approaches (e.g., KL divergence)

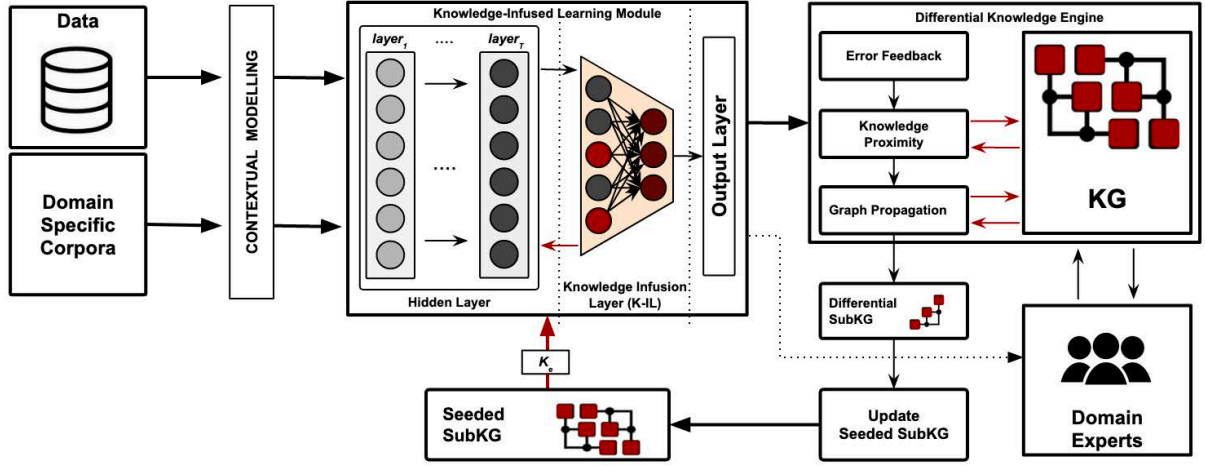


Figure 2: Overall Architecture: Contextual representations of data are generated, and domain knowledge amplifies the significance of specific important concepts that are missed in the learning model. Classification error determines the need for updating a Seeded SubKG with more relevant knowledge, resulting in a Seeded SubKG that is more refined and informative to our model.

and probabilistic soft logic (Kimmig et al. 2012). Further, a sub-graph discovery approach (Cameron et al. 2015; Lalithsena 2018) can also be used utilizing probabilistic graphical models (e.g., deep belief networks, conditional random fields). In our approach, the Seeded SubKG will be updated with more knowledge based on difference between the learned representation and relevant knowledge representation from the KG (see Section Differential Knowledge Engine).

K_e : Knowledge Embedding Creation Representation of knowledge in the Seeded SubKG will be generated as embedding vectors. Specific contextual dimension models and/or more generic models can be utilized to create an embedding of each concept and their relations in the Seeded SubKG. Unlike traditional approaches that compute the representation of each concept in the KGs by simply taking an average of embedding vectors of concepts, we leverage the existing structural information of the graph. This procedure is formally defined:

$$K_e = \sum_{ij} [C_i, C_j] \otimes D_{ij} \quad (1)$$

where K_e is the representation of the concepts enriched by the relationships in the Seeded-KG, (C_i, C_j) is the relevant pair of concepts in the Seeded-KG, D_{ij} is the distance measure (e.g., Least Common Subsumer (Baader, Sertkaya, and Turhan 2007)) between the two concepts C_i and C_j . Novel methods will further be examined building upon this initial approach above as well as existing tools that include TRANS-E (Bordes et al. 2013), TRANS-H (Wang et al. 2014), and HOLE (Nickel et al. 2016) for the creation of embeddings from KGs.

Knowledge Infusion Layer In a many-to-one NLM (Shivakumar et al. 2018) network with T hidden layers, the T^{th} layer contains the learned representation before the

output layer. The output layer (e.g., SoftMax) of the NLM model estimates the error to be back-propagated. As the techniques for knowledge infusion between hidden layers or just before the output layers will be explored, in this subsection, we explain the Knowledge Infusion Layer (K-IL) which takes place just before the output layer.

Algorithm 1 Routine for Infusion of Knowledge in NLMs

```

1: procedure KNOWLEDGEINFUSION
2:   Data :  $NLM_{type}, \#Epochs, \#Iter, K_e$ 
3:   Output :  $\vec{M}_T$ 
4:   for  $ne=1$  to  $\#Epochs$  do
5:      $\vec{h}_T, \vec{h}_{T-1} \leftarrow \text{TrainingNLM}(NLM_{type}, \#Iter)$ 
6:     while  $(\mathbf{D}_{KL}(\vec{h}_{T-1} || \vec{K}_e) - \mathbf{D}_{KL}(\vec{h}_T || \vec{K}_e)) > \epsilon$  do
7:        $\vec{h}_T \leftarrow \sigma(W_{hk} * (\vec{h}_T \oplus \vec{K}_e) + b_{hk})$ 
8:        $W^{hk} \leftarrow W^{hk} - \eta_k \nabla(\mathbf{K-LF})$ 
9:        $\vec{M}_T \leftarrow \vec{h}_T \odot W^{hk}$ 
10:    return  $\vec{M}_T$ 

```

Algorithm 1 takes the type of neural language model, number of epochs, iterations and the seeded knowledge graph embedding K_e as input, and returns a knowledge infused representation of the hidden state \vec{M}_T . In line 4, the infusion of knowledge takes place after each epoch without obstructing the learning of the vanilla NLM model and is explained in lines 5-10. Within the knowledge infusion process (lines 7-9), we optimize the loss function in equation 2 with convergence condition defined as the reduction in the difference between the \mathbf{D}_{KL} of \vec{h}_T and \vec{h}_{T-1} in the presence of K_e . Considering the vanilla structure of a NLM (Greff et al. 2017), \vec{M}_T is utilized by the fully connected layer for classification.

To illustrate an initial approach in Figure 3, we use LSTMs as NLMs in our neural network. K-IL functions add an additional layer before the output layer of our proposed

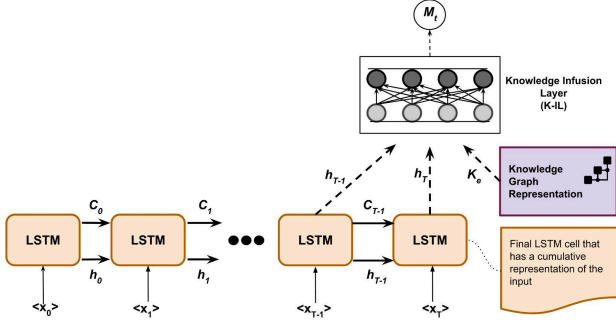


Figure 3: Inner Mechanism of the Knowledge Infusion Layer

neural network architecture. This layer takes the latent vector (\mathbf{h}_{T-1}) of the penultimate layer, the latent vector of the last hidden layer (\mathbf{h}_T) and the knowledge embedding (K_e), as input.

In this layer, we define two particular functions that will be critical for merging the latent vectors from the hidden layers and the knowledge embedding vector from the KG. Note that the dimensions of these vectors are the same because they are created from the same models (e.g., contextual models), which makes the merge operation of those vectors possible and valid.

K-LF: Knowledge-Aware Loss Function In neural networks, hidden layers may de-emphasize important patterns due to the sparsity of certain features during learning, which causes information loss. In some cases, such patterns may not even appear in the data. However, such relations or patterns may be defined in KGs with even more relevant knowledge. We call this information gap between the learned representation of the data and knowledge representation as **differential knowledge**. Information loss in a learning process is relative to the distribution that suffered the loss. Hence, we propose a measure to determine the *differential knowledge* and guide the degree of knowledge infusion in learning. As our initial approach to this measure, we developed a two-state regularized loss function by utilizing Kullback Leibler (KL) divergence. Our choice of KL divergence measure is largely influenced by the Markov assumptions made in language modeling and have been highlighted in (Longworth 2010). The K-LF measure estimates the divergence between the hidden representations (\mathbf{h}_{T-1} , \mathbf{h}_T) and knowledge representation (K_e), to determine the differential knowledge to be infused.

Formally we define it as:

$\arg \min(h_{T-1}, h_T, K_e) \equiv \text{K-LF}$, where \mathbf{h}_{T-1} is an input for convergence constraint.

$$\begin{aligned} \text{K-LF} &= \min \mathbf{D}_{KL}(h_T || K_e); \\ \text{s.t. } \mathbf{D}_{KL}(h_T || K_e) &< \mathbf{D}_{KL}(h_{T-1} || K_e) \end{aligned} \quad (2)$$

We minimize the *relative entropy* for information loss to

maximize the information gain from the knowledge representation (e.g., K_e). We will compute differential knowledge ($\nabla \text{K-LF}$) through such optimization approach; thus, the computed differential knowledge will also determine the degree of knowledge to be infused. $\nabla \text{K-LF}$ will be computed in the form of embedding vectors, and the dimensions from K_e will be preserved.

K-MF: Knowledge Modulation Function We need to merge the differential knowledge representation with the partially learned representation. However, this operation cannot be done arbitrarily as the vector spaces of both representations are different both in dimension and distribution if not same (Dumančić and Blockeel 2017). We explain an initial approach for the K-MF to modulate the learned weight matrix of the neural network with the hidden vector through an appropriate operation (e.g., Hadamard pointwise multiplication). This operation at the T^{th} layer can be formulated as:

Equation for $W^{hk} = W^{hk} - \eta_k * \nabla \text{K-LF}$, where W^{hk} is the learned weight matrix infusing knowledge, η_k is learning momentum (Sutskever et al. 2013), $\nabla \text{K-LF}$ is differential knowledge. The weight matrix (W^{hk}) is computed through the learning epochs utilizing the differential knowledge embedding ($\nabla \text{K-LF}$). Then we merge W^{hk} with the hidden vector \mathbf{h}_T through the K-MF. Considering that we use Hadamard pointwise multiplication as our initial approach, we formally define the output \mathbf{M}_T of K-MF as:

This operation at the T^{th} layer can be formulated as:

$$\vec{M}_T = \vec{h}_T \odot W^{hk} \quad (3)$$

where \mathbf{M}_T is Knowledge-Modulated representation, \mathbf{h}_T is the hidden vector and W^{hk} is the learned weight matrix infusing knowledge. Further investigations of techniques for K-MF constitutes a central research topic for the research community.

Differential Knowledge Engine

In deep neural networks, each epoch generates an error that is back-propagated until the model reaches a saddle point in the local minima, and the error is reduced in each epoch. The error indicates the difference between probabilities of actual and predicted labels, and this difference can be used to enrich the Seeded SubKG in our proposed knowledge-infused learning (K-IL) framework.

In this subsection, we discuss the sub-knowledge graph operations that are based on the difference between the learned representation of our knowledge-infused model (\mathbf{M}_T), and the representation of the relevant sub-knowledge graph from the KG, which we name the differential sub-knowledge graph. We define a *Knowledge Proximity function* to generate the *Differential Sub-knowledge Graph*, and *Update Seeded SubKG* to insert the differential sub-knowledge graph into the Seeded SubKG.

Knowledge Proximity Upon the arrival of the learned representation from the knowledge-infused learning model, we query the KG for retrieving related information to the respective data point. In this particular step, it is important

to find the optimal proximity between the concept and its related concepts. For example, from the “South Carolina” concept, we may traverse the surrounding concepts with a varying number of hops (empirically decided). Finding the optimal number of hops towards each direction from the concept in question is still an open research question. As we find optimal proximity of a particular concept in the KG, we propagate KG based on the proximity starting from the concept in question.

Differential SubKG Once we obtain the SubKG from the graph propagation, we create a differential SubKG that will reflect the difference in knowledge from the Seeded SubKG. For this procedure, research is needed to formulate the problem using variational autoencoders to extract a *differential subKG*(D_{kg}) and, we believe it will provide missing information in the Seeded-KG.

Update function The differential subKG generated as a result of minimizing knowledge proximity is considered as an input factual graph to the update procedure. As a result, the procedure dynamically evolves the Seeded subKG with missing information from differential subKG. We propose to utilize *Lyapunov stability theorem* (Liu, Zhang, and Chen 2014) and *Zero Shot learning* to update the Seeded-KG using D_{kg} . D_{kg} and Seeded-KG represent two knowledge structures requiring a process of transferring the knowledge from one structure to another (Hamaguchi et al. 2017). We define this process as generation of semantic mapping weights that encodes and decodes the two semantic spaces, utilizing the Lyapunov stability constraint and Sylvester optimization approach: Given two semantic spaces belonging to a domain D (e.g., online extremism, mental health), we tend to attain an equilibrium position defined as:

$$\|S_{kg} - W * D_{kg}\|_F^2 = \alpha * \|W * S_{kg} - D_{kg}\|_F^2 \quad (4)$$

$\|\cdot\|_F$ represents Frobenius norm and α is a proportionality constant belong to \mathbb{R} . Equation 4 reflects Lyapunov stability theorem and to achieve such a stable state we define our optimization function as follows:

$$L = \min(\|S_{kg} - W D_{kg}\|_F^2 - \alpha * \|W S_{kg} - D_{kg}\|_F^2), \\ \alpha > 0, W \in \mathbb{R}^{X \times \mathbb{R}} \quad (5)$$

Equation 5 is solvable using Sylvester optimization and its derivation is defined in a recent study (Gaur et al. 2018).

Applications for K-IL

Artificial intelligence models will be widely deployed in real world decision making processes in the foreseeable future, once the challenges described in Section 1, are overcome. As we argue that the incorporation of external structured knowledge will address these challenges, it will benefit various application domains such as social and health sciences, automating processes that require knowledge and intelligence. Specifically, it will have a potentially significant impact on

predictive analysis of online communications such as misinformation and extremism, conversational modeling, and disease prediction.

As predicting online extremism is challenging and false alarms create serious implications potentially affecting millions of individuals, (Kursuncu et al. 2019a) showcased that the (shallow) infusion of external domain-specific knowledge improves precision, reducing potential social discrimination. Further, in prediction of mental health diseases defined in DSM-5, (Gaur et al. 2018) shallow knowledge infusion reduces false alarms by 30%. On the other hand, conversational models pose an important application area as (Liu et al. 2019b) proposed a conversation framework where the fusion of KGs and text mutually reinforce each other to generate knowledge-aware responses, improving the model in generalizability and explainability. In another study, (Young et al. 2018) integrated commonsense knowledge into the conversational models selecting the most appropriate response. While machine learning finds many application areas in medicine for disease prediction, large data is not always available. In this case knowledge-infused learning generates more representative features thereby avoiding overfitting. A study (Tan et al. 2019) on early diagnosis of lung cancer using computed tomography images, infused knowledge in the form of expert-curated features into the learning process through CNN. Despite the small data set, the enriched feature space in their knowledge-infused learning process improved sensitivity and specificity of the model.

In contrast to the applications above, we believe that the deep infusion of external knowledge within latent layers will enhance the coverage of the information being learned by the model based on KGs. Hence, this will provide better generalizability, reduction in bias and false alarms, disambiguation, less reliance on large data, explainability, reliability and robustness, to the real world applications in critical aforementioned domains with significant impact.

Conclusion

Combining deep learning and knowledge graphs in a hybrid neural-symbolic learning framework will further enhance performance and accelerate the convergence of the learning processes. Specifically, the impact of this improvement in very sensitive domains such as health and social science, will be significant with respect to their implications for real-world deployment. Adoption of the tools that automate tasks that require knowledge and intelligence, and are traditionally done by humans, will improve with the help of this framework that marries deep learning and knowledge graph techniques. Specifically, we envision that the infusion of knowledge as described in this framework will capture information for the corresponding domain in finer granularity of abstraction. We believe that this approach will provide reliable solutions to the problems faced in deep learning, as described in Sections 1 and 5. Hence, in real world applications, resolving these issues with both knowledge graphs and deep learning in a hybrid neuro-symbolic framework will greatly contribute to fulfilling AI’s promise.

Acknowledgement

We acknowledge partial support from the National Science Foundation (NSF) award CNS-1513721: "Context-Aware Harassment Detection on Social Media". Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

- Athiwaratkun, B.; Wilson, A. G.; and Anandkumar, A. 2018. Probabilistic fasttext for multi-sense word embeddings. *arXiv preprint arXiv:1806.02901*.
- Baader, F.; Sertkaya, B.; and Turhan, A.-Y. 2007. Computing the least common subsumer wrt a background terminology. *Journal of Applied Logic* 5(3):392–420.
- Bhatt, S.; Gaur, M.; Bullemer, B.; Shalin, V.; Sheth, A.; and Minnery, B. 2018a. Enhancing crowd wisdom using explainable diversity inferred from social media. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 293–300. IEEE.
- Bhatt, S.; Gaur, M.; Bullemer, B.; Shalin, V. L.; Sheth, A. P.; and Minnery, B. 2018b. Enhancing crowd wisdom using explainable diversity inferred from social media. In *IEEE/WIC/ACM International Conference on Web Intelligence*. Santiago, Chile: IEEE.
- Bian, J.; Gao, B.; and Liu, T.-Y. 2014. Knowledge-powered deep learning for word embedding. In *Joint European conference on machine learning and knowledge discovery in databases*, 132–148. Springer.
- Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; and Hellmann, S. 2009. Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web* 7(3):154–165.
- Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250. AcM.
- Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, 2787–2795.
- Brouch, K. L. 2000. Where in the world is icd-10? *Where in the World Is ICD-10?/AHIMA, American Health Information Management Association*.
- Cameron, D.; Kavuluru, R.; Rindflesch, T. C.; Sheth, A. P.; Thirunarayan, K.; and Bodenreider, O. 2015. Context-driven automatic subgraph creation for literature-based discovery. *Journal of biomedical informatics* 54:141–157.
- Casteleiro, M. A.; Demetriou, G.; Read, W.; Prieto, M. J. F.; Maroto, N.; Fernandez, D. M.; Nenadic, G.; Klein, J.; Keane, J.; and Stevens, R. 2018. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. *Journal of biomedical semantics* 9(1):13.
- Chen, B.; Hao, Z.; Cai, X.; Cai, R.; Wen, W.; Zhu, J.; and Xie, G. 2019. Embedding logic rules into recurrent neural networks. *IEEE Access* 7:14938–14946.
- Cheng, J. 2018. Ai reasoning systems: Pac and applied methods. *arXiv preprint arXiv:1807.05054*.
- Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Corbitt-Hall, D. J.; Gauthier, J. M.; Davis, M. T.; and Witte, T. K. 2016. College students' responses to suicidal content on social networking sites: an examination using a simulated facebook newsfeed. *Suicide and Life-Threatening Behavior* 46(5):609–624.
- De Palma, G.; Kiani, B.; and Lloyd, S. 2019. Random deep neural networks are biased towards simple functions. In *Advances in Neural Information Processing Systems*, 1962–1974.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dugas, C.; Bengio, Y.; Bélisle, F.; Nadeau, C.; and Garcia, R. 2009. Incorporating functional knowledge in neural networks. *Journal of Machine Learning Research* 10(Jun):1239–1262.
- Dumančić, S., and Blockeel, H. 2017. Demystifying relational latent representations. In *International Conference on Inductive Logic Programming*, 63–77. Springer.
- Gaur, M.; Kursuncu, U.; Alambo, A.; Sheth, A.; Daniulaityte, R.; Thirunarayan, K.; and Pathak, J. 2018. "let me tell you about your mental health!" contextualized classification of reddit posts to dsm-5 for web-based intervention.
- Gaur, M.; Alambo, A.; Sain, J. P.; Kursuncu, U.; Thirunarayan, K.; Kavuluru, R.; Sheth, A.; Welton, R.; and Pathak, J. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, 514–525.
- Goldberg, Y., and Levy, O. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.
- Greff, K.; Srivastava, R. K.; Koutník, J.; Steunebrink, B. R.; and Schmidhuber, J. 2017. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28(10):2222–2232.
- Gruber, T. 2008. Ontology, encyclopedia of database systems, ling liu and m. Tamer Özsu (Eds.).
- Halevy, A.; Norvig, P.; and Pereira, F. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24(2):8–12.
- Hamaguchi, T.; Oiwa, H.; Shimbo, M.; and Matsumoto, Y. 2017. Knowledge transfer for out-of-knowledge-base entities: a graph neural network approach. *arXiv preprint arXiv:1706.05674*.
- Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

- Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- Hoffart, J.; Suchanek, F. M.; Berberich, K.; and Weikum, G. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence* 194:28–61.
- Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; and Xing, E. 2016. Harnessing deep neural networks with logic rules. *arXiv preprint arXiv:1603.06318*.
- Islam, S. R.; Eberle, W.; Bundy, S.; and Ghafoor, S. K. 2019. Infusing domain knowledge in ai-based "black box" models for better explainability with application in bankruptcy prediction. *arXiv preprint arXiv:1905.11474*.
- Karpathy, A. 2015. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog* 21.
- Kho, S. J.; Padhee, S.; Bajaj, G.; Thirunarayan, K.; and Sheth, A. 2019. Domain-specific use cases for knowledge-enabled social media analysis. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer. 233–246.
- Kimmig, A.; Bach, S.; Broecheler, M.; Huang, B.; and Getoor, L. 2012. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 1–4.
- Kursuncu, U.; Gaur, M.; Lokala, U.; Illendula, A.; Thirunarayan, K.; Daniulaityte, R.; Sheth, A.; and Arpinar, I. B. 2018. "what's ur type?" contextualized classification of user types in marijuana-related communications using compositional multiview embedding. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'18)*.
- Kursuncu, U.; Gaur, M.; Castillo, C.; Alambo, A.; Thirunarayan, K.; Shalin, V.; Achilov, D.; Arpinar, I. B.; and Sheth, A. 2019a. Modeling islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):151.
- Kursuncu, U.; Gaur, M.; Lokala, U.; Thirunarayan, K.; Sheth, A.; and Arpinar, I. B. 2019b. Predictive analysis on twitter: Techniques and applications. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer. 67–104.
- Kursuncu, U.; Gaur, M.; Thirunarayan, K.; and Sheth, A. 2019c. Explainability of medical ai through domain knowledge. *Ontology Summit 2019, Medical Explanation*.
- Kursuncu, U. 2018. *Modeling the Persona in Persuasive Discourse on Social Media Using Context-aware and Knowledge-driven Learning*. Ph.D. Dissertation, University of Georgia.
- Lalithsena, S. 2018. *Domain-specific knowledge extraction from the web of data*. Ph.D. Dissertation, Wright State University.
- Liu, H., and Singh, P. 2004. Commonsense reasoning in and over natural language. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 293–306. Springer.
- Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2019a. K-bert: Enabling language representation with knowledge graph. *arXiv preprint arXiv:1909.07606*.
- Liu, Z.; Niu, Z.-Y.; Wu, H.; and Wang, H. 2019b. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1782–1792.
- Liu, M.; Zhang, D.; and Chen, S. 2014. Attribute relation learning for zero-shot classification. *Neurocomputing* 139:34–46.
- Longworth, C. 2010. *Kernel methods for text-independent speaker verification*. Ph.D. Dissertation, University of Cambridge.
- Makni, B., and Hendler, J. Deep learning for noise-tolerant rdfs reasoning.
- Masse, N. Y.; Grant, G. D.; and Freedman, D. J. 2018. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences* 115(44):E10467–E10475.
- Maurya, A. K. 2018. Learning low dimensional word based linear classifiers using data shared adaptive bootstrap aggregated lasso with application to imdb data. *arXiv preprint arXiv:1807.10623*.
- McInnes, B. T.; Pedersen, T.; and Pakhomov, S. V. 2009. Umls-interface and umls-similarity: open source software for measuring paths and semantic similarity. In *AMIA Annual Symposium Proceedings*, volume 2009, 431. American Medical Informatics Association.
- Nickel, M.; Rosasco, L.; Poggio, T. A.; et al. 2016. Holographic embeddings of knowledge graphs. In *AAAI*, volume 2, 3–2.
- Ohno-Machado, L.; Sansone, S.-A.; Alter, G.; Fore, I.; Grethe, J.; Xu, H.; Gonzalez-Beltran, A.; Rocca-Serra, P.; Gururaj, A. E.; Bell, E.; et al. 2017. Finding useful data across multiple biomedical data repositories using datamed. *Nature genetics* 49(6):816.
- Olteanu, A.; Castillo, C.; Diaz, F.; and Kiciman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2:13.
- Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, 1410–1418.
- Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.
- Perera, S.; Mendes, P. N.; Alex, A.; Sheth, A. P.; and Thirunarayan, K. 2016. Implicit entity linking in tweets. In *International Semantic Web Conference*, 118–132. Springer.
- Roy, A.; Park, Y.; and Pan, S. 2017. Learning domain-specific word embeddings from sparse cybersecurity texts. *arXiv preprint arXiv:1709.07470*.

- Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206–215.
- Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.
- Sarker, M. K.; Xie, N.; Doran, D.; Raymer, M.; and Hitzler, P. 2017. Explaining trained neural networks with semantic web technologies: First steps. *arXiv preprint arXiv:1710.04324*.
- Scarlino, B.; Pasini, T.; and Navigli, R. 2020. SenseBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proc. of AAAI*.
- Shen, Y.; Deng, Y.; Yang, M.; Li, Y.; Du, N.; Fan, W.; and Lei, K. 2018. Knowledge-aware attentive neural network for ranking question answer pairs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 901–904. ACM.
- Sheth, A., and Kapanipathi, P. 2016. Semantic filtering for social data. *IEEE Internet Computing* 20(4):74–78.
- Sheth, A., and Thirunarayan, K. 2012. Semantics empowered web 3.0: managing enterprise, social, sensor, and cloud-based data and services for advanced applications. *Synthesis Lectures on Data Management* 4(6):1–175.
- Sheth, A.; Perera, S.; Wijeratne, S.; and Thirunarayan, K. 2017. Knowledge will propel machine understanding of content: Extrapolating from current examples. *arXiv preprint arXiv:1707.05308*.
- Sheth, A.; Gaur, M.; Kursuncu, U.; and Wickramarachchi, R. 2019. Shades of knowledge-infused learning for enhancing deep learning. *IEEE Internet Computing* 23(6):54–63.
- Shivakumar, P. G.; Li, H.; Knight, K.; and Georgiou, P. 2018. Learning from past mistakes: Improving automatic speech recognition output via noisy-clean phrase context modeling. *arXiv preprint arXiv:1802.02607*.
- Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 843–852. IEEE.
- Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.
- Sutskever, I.; Martens, J.; Dahl, G.; and Hinton, G. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, 1139–1147.
- Tan, J.; Huo, Y.; Liang, Z.; and Li, L. 2019. Expert knowledge-infused deep learning for automatic lung nodule detection. *Journal of X-ray science and technology* 27(1):17–35.
- Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25(1):44–56.
- Valiant, L. G. 2000. Robust logics. *Artificial Intelligence* 117(2):231–253.
- Vo, K.; Pham, D.; Nguyen, M.; Mai, T.; and Quan, T. 2017. Combination of domain knowledge and deep learning for sentiment analysis. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, 162–173. Springer.
- Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, volume 14, 1112–1119.
- Yi, K.; Jian, Z.; Chen, S.; Chen, Y.; and Zheng, N. 2018. Knowledge-based recurrent attentive neural network for traffic sign detection. *arXiv preprint arXiv:1803.05263*.
- Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.