Leveraging Symbolic Knowledge Bases for Commonsense Natural Language Inference using Pattern Theory

Sathyanarayanan N. Aakur, *Member, IEEE,* Sudeep Sarkar, *Fellow, IEEE.*

Abstract—The commonsense natural language inference (CNLI) tasks aim to select the most likely follow-up statement to a contextual description of ordinary, everyday events and facts. Current approaches to transfer learning of CNLI models across tasks require many labeled data from the new task. This paper presents a way to reduce this need for additional annotated training data from the new task by leveraging symbolic knowledge bases, such as ConceptNet. We formulate a teacher-student framework for mixed symbolic-neural reasoning, with the large-scale symbolic knowledge base serving as the teacher and a trained CNLI model as the student. This hybrid distillation process involves two steps. The first step is a symbolic reasoning process. Given a collection of unlabeled data, we use an abductive reasoning framework based on Grenander's pattern theory to create weakly labeled data. Pattern theory is an energy-based graphical probabilistic framework for reasoning among random variables with varying dependency structures. In the second step, the weakly labeled data, along with a fraction of the labeled data, is used to transfer-learn the CNLI model into the new task. The goal is to reduce the fraction of labeled data required. We demonstrate the efficacy of our approach by using three publicly available datasets (OpenBookQA, SWAG, and HellaSWAG) and evaluating three CNLI models (BERT, LSTM, and ESIM) that represent different tasks. We show that, on average, we achieve 63% of the top performance of a fully supervised BERT model with no labeled data. With only 1000 labeled samples, we can improve this performance to 72%. Interestingly, without training, the teacher mechanism itself has significant inference power. The pattern theory framework achieves 32.7% accuracy on OpenBookQA, outperforming transformer-based models such as GPT (26.6%), GPT-2 (30.2%), and BERT (27.1%) by a significant margin. We demonstrate that the framework can be generalized to successfully train neural CNLI models using knowledge distillation under unsupervised and semi-supervised learning settings. Our results show that it outperforms all unsupervised and weakly supervised baselines and some early supervised approaches, while offering competitive performance with fully supervised baselines. Additionally, we show that the abductive learning framework can be adapted for other downstream tasks, such as unsupervised semantic textual similarity, unsupervised sentiment classification, and zero-shot text classification, without significant modification to the framework. Finally, user studies show that the generated interpretations enhance its explainability by providing key insights into its reasoning mechanism.

Index Terms—Multiple choice CNLI, Commonsense Reasoning, Pattern Theory

1 Introduction

We can partition natural language understanding into different problem domains, such as classification, commonsense reasoning, machine reading comprehension, and summarization. Within each domain, there are various specific tasks. One open problem is task transfer learning, which involves transferring a model from a source task to a different target task within a specific domain. Typical solutions require a large amount of labeled data from the target domain. However, we consider task transfer learning with the constraint that we have only a minimal set of labeled data in the target domain but have access to a symbolic commonsense knowledge base. Although the underlying problem formulation, i.e., text classification, may be similar, each task presents different challenges, such as domain-specific semantics, multi-hop reasoning, and contextual information, that make them distinct from one another. For example, answering questions about everyday events requires a different set of reasoning capabilities than answering open-domain, fact-based questions, even though both tasks fall under the broader category of question-answering. In this work, we focus primarily on the different commonsense natural language inference (CNLI) tasks in the commonsense reasoning domain.

Common formulations of CNLI tasks involve selecting the most likely follow-up statement from a list of choices in specific domains such as everyday facts and events. For instance, the SWAG task (Situations With Adversarial Generations) [1] consists of multiple-choice sentence completions derived from captions of consecutive events of videos in ActivityNet [2] and the Large Scale Movie Description Challenge (LSMDC) [3]. The questions span many domains, so formulating a complete solution requires reasoning over prior knowledge, establishing semantic relationships among entities, and language comprehension. Other examples of CNLI tasks include the general-purpose knowledge inference task (OpenBookQA [4]) and the how-to instruction tasks (HellaSWAG [5]). The typical approach to solving such tasks has been to adapt pre-trained deep-networks-based language models such as BERT [6], GPT [7], and more recently, GPT-3 [8] for a specific CNLI task in a supervised manner. This approach has yielded state-of-

- *SN Aakur is with the Department of Computer Science, Oklahoma State University, Stillwater, OK, 74078. E-mail: saakurn@okstate.edu*
- *Sudeep Sarkar is with the Department of Computer Science and Engineering at the University of South Florida, Tampa, FL, 33620. E-mail: sarkar@usf.edu*

Manuscript received April 19, 2005; revised August 26, 2015.

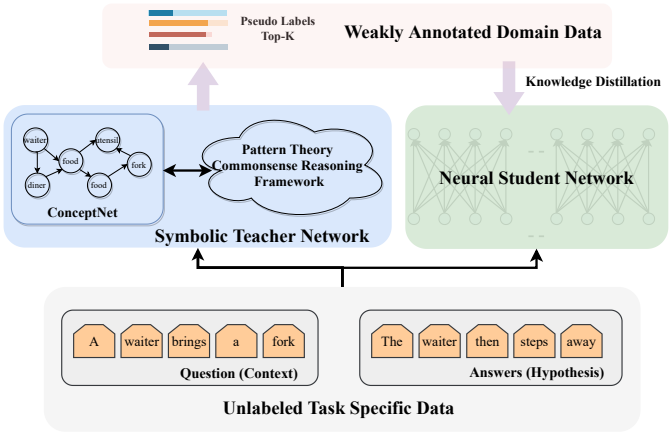

Fig. 1. Given unlabeled data from the target task, we use a general commonsense reasoning framework based on pattern theory to create weakly-labeled data using symbolic knowledge bases, e.g., ConceptNet. This is the teacher. We then distill its implicit knowledge to train a student to build specialist models for the target task.

the-art performance on several CNLI benchmarks through large-scale pre-training on vast amounts of unlabeled data.

Despite such success of pre-trained models, switching the trained model from one CNLI task (the source) to another (the target) is a harder problem. As an illustration, consider the BERT model for three CNLI tasks: (i) everyday situations (SWAG) [1], (ii) general-purpose knowledge [4] (OpenBookQA), and (iii) how-to instructions (HellaSWAG) [5]. To switch a BERT model trained on, say, SWAG (the source task) to OpenBookQA (the target) requires the availability of large amounts of *labeled* training data in the target task, i.e., OpenBookQA. Table 1 shows the transferability of BERT across tasks with different amounts of labeled data in the target tasks: (i) when a large amount of *labeled* training data is available for the target task, (ii) when only $1,000$ *labeled* training samples are available in the target task, and (iii) when no *labeled* training data is available at all. The highlighted diagonal percentages represent the best BERT performance that can be achieved on the respective task. The off-diagonal percentages represent the generalization performance from the source to the target task. We observe a significant drop in performance for all cross-task generalization scenarios. The performance is poor across tasks when there is no labeled data. With a large amount of labeled data, the performance is better, but it drops when there is limited labeled data available for the target task. Training with robust adversarial filtering seems to reduce overfitting and helps models trained on HellaSWAG to generalize to out-of-task data. However, it requires careful selection and refinement of training data.

To reduce the dependence on labeled training data, we utilize the knowledge stored in large-scale symbolic knowledge bases such as ConceptNet [9], [10] to provide weak supervision for CNLI models in the target task. The approach builds on the idea of *abductive reasoning* [11] for distilling knowledge from the symbolic knowledge base. Figure 1 shows the overall approach. We start with unlabeled data in the target domain and generate weakly labeled data using a pattern theory-based reasoning framework that leverages large-scale symbolic knowledge bases. For training CNLI models, we use a student-teacher setup introduced in knowledge distillation [12], [13]. However, unlike the standard teacher-student setup where both the teacher and the student are deep-learning models, we have a hybrid setup where the teacher is the large-scale symbolic knowledge base, and the student is the CNLI model to be trained. We formulate the teacher-student distillation as a two-step process. First, we use a pattern theory-based inference engine to weakly label the data by leveraging the commonsense knowledge base. Second, we use this weakly-labeled data along with an optional fraction of *labeled* training data to train CNLI models on task-specific data. Our approach differs from works such as COMET [14], which uses large-scale information in deep neural networks for knowledge base expansion and completion. Instead, we use large-scale knowledge to develop task-specific models.

We use Grenander's pattern theory formalism [15] to express this reasoning framework. Pattern theory is a graphical, energy-based probabilistic framework that can reason over random variables with varying dependency structures. The underlying structure is represented as compositions of simpler patterns. Each element of the structure, called generators, combines with each other through local interactions via links called bonds. These interactions are constrained by both local and global regularities captured by an overarching graph structure. A probability structure over the representations captures the diversity of patterns. The many incarnations of graphical models of patterns, such as directed acyclic graphs (DAG), Markov random fields (MRF), Gaussian random fields, and formal languages, can be shown to be special cases (see Chapter 6 of [16]).

A significant departure from current approaches to CNLI is the use of symbolic reasoning to first construct a "*contextualized interpretation*" of the evidence (the question or context) and each of the provided hypotheses (the answer choices), expressed in a graph-like structure using pattern theory. An example of a contextualized interpretation is illustrated in Figure 4. We define an interpretation as a connected representation that captures the semantic structure of the evidence. An interpretation is a deeper and more meaningful representation of observed concepts (actors, actions, and actor-object interactions) and unobserved concepts (background knowledge of concepts) or "contextualization cues." We use these interpretations to perform "*inference to the best explanation*" (IBE) to find the most plausible hypothesis.

To demonstrate the effectiveness of the proposed framework, we chose unsupervised commonsense natural language inference as the primary task for evaluation. The CNLI task is naturally conducive to abductive reasoning since it requires reasoning over observations in the context of prior knowledge to ascertain plausibility. It requires complex, multi-hop reasoning that goes beyond simple pairwise relationships and requires a deeper understanding of the semantic relationships among concepts in the hypotheses, especially in an unsupervised setting without gold-standard labels. We show that the framework can be expanded, without significant rewiring, into other downstream tasks, such as semantic textual similarity [17] (Section 6.3), sentiment analysis [18] (Section 6.4), and zero-shot text classification (Section 6.5), while providing an explainable interface (Section 6.2) to the underlying reasoning mechanism.

TABLE 1
Transferability of BERT across CNLI tasks: **O**penBookQA (OBQA), **S**WAG and **H**ellaSWAG (HSWAG). We list the accuracies for different combinations of source and target tasks. We evaluate performance under three different sizes of the **labeled** training data from the target task: (i) the entire labeled set is available, (ii) when only $1,000$ labeled samples are available, and (c) when no labeled sample is available.

Source task	Target task						
	All Labeled Samples			1000 Labeled Samples			None
	OBQA	SWAG	HSWAG	OBQA	SWAG	HSWAG	
OBQA	**56.6**	30.6	30.2	**38.8**	24.4	24.6	27.1
SWAG	20.03	**86.6**	71.4	28.2	**58.5**	41.9	**41.4**
HSWAG	25.63	34.6	**46.7**	26.1	26.7	**29.8**	28.9

The **contributions** of this work are that we have formulated a novel *pattern theory-based abductive reasoning* framework to abstract task-relevant information in large-scale symbolic knowledge bases into task-specific neural networks. This hybrid knowledge distillation mechanism is new and can be used to train CNLI models using large-scale symbolic knowledge bases with few labeled training data.

We have structured the paper as follows: In Section 2, we review related work on the methods and techniques used in our work. The overview of the approach is outlined in Section 3, followed by details of Grenander's pattern theory-based formulation of the symbolic teacher in Section 4. In Section 5, we show how the knowledge is distilled into the task-specific student network. Sections 6.2, 6.3, 6.4, 6.5, and 6.6 present a thorough performance evaluation of the proposed approach along with ablation studies. Section 7 provides error analysis and discusses future directions for error mitigation.

2 Related Work

Commonsense natural language inference (CNLI) has primarily been addressed in current literature as a type of question-answering, along with other tasks such as comprehension [19] and natural language inference (NLI) [1], [5], [20]. Related downstream tasks include fact-checking [21] and semantic textual similarity [17], [22], which use CNLI to assess the factual and semantic accuracy of text. Approaches to these tasks can be divided into two categories: semantic similarity matching and relevance matching models. Similarity matching models involve computing semantic similarity between question and answer representations, typically using a neural network model such as BERT [6], OpenAI GPT [7], ESIM [23] and Fast-Text [24] and LSTM based approaches. Other approaches use a "compare, attend, and aggregate" framework to quantify the relevance between answers and questions [25], starting with vector representations of both and aggregating the relevance for a final prediction. Other approaches represent some of the early supervised models such as FastText [24], which use a bag of words to represent the language for QA.

Commonsense knowledge bases are large repositories of structured knowledge extracted from raw textual data that express relational information between entities present in everyday facts and events. They are typically represented as graphs or hypergraphs, with nodes consisting of concepts and edges expressing the relationship between them. Over the years, several knowledge bases have been curated, such as ConceptNet [10], Cyc [26], FrameNet [27], DBPedia [28], WordNet [29], and ATOMIC [30], each focusing on capturing a specific aspect of commonsense knowledge. For example, ConceptNet captures the semantic relationships between concepts through a hypergraph, with edges spanning 34 different assertions such as `IsA`, `RelatedTo`, `AtLocation`, and more. ATOMIC focuses on inferential knowledge, capturing 9 if-then relations expressed over variables to encode cause-vs-effect and agent-vs-theme knowledge. The knowledge expressed in these large-scale repositories is typically manually curated, with recent efforts focusing on knowledge base completion [31], [32] to expand existing knowledge bases by predicting relationships between concepts. While previous work has focused on supervised learning to leverage knowledge bases for various tasks in natural language processing and computer vision [33], [34], [35], our student-teacher framework eliminates the need for supervised training by using the inherent symbolic knowledge in large-scale knowledge bases as the teacher to distill commonsense knowledge and train student models for downstream tasks such as CNLI.

Knowledge-based approaches to question answering [36], [37], [38], [39] have gained traction to reduce the increasing reliance on large, human-annotated datasets for commonsense NLI. Such approaches construct large repositories of knowledge by enhancing existing sources of knowledge, such as ConceptNet [10] and ATOMIC [30], with auxiliary, domain-specific knowledge extracted from text, such as QASC [40]. Synthetic question-answer pairs are constructed from these custom-built knowledge bases to pretrain language models for zero-shot and few-shot question answering. Some approaches, such as KagNet [41], KTL [37], MHGRN [42], QAGNN [38], OCN [43], KEAR [44], and KnowledgePath [39], to name a few, have integrated commonsense knowledge found in symbolic knowledge bases, such as ConceptNet and ATOMIC, into neural networks using knowledge-injection techniques (such as attention and graph neural networks) to enhance performance on CNLI tasks through supervised learning. Other approaches leverage the knowledge captured in large language models, such as BERT [6], as supervision for CNLI using different mechanisms, such as consistency optimization [45], question rewriting [46], and leveraging the autoregressive pretraining objective to rank answer options [47], [48]. Our approach falls under this category of models that reduce the requirements for annotated training data for commonsense NLI. However, we do not require the construction of additional, specialized knowledge bases, additional mechanisms for question rewriting, or ensembling for CNLI. Further-

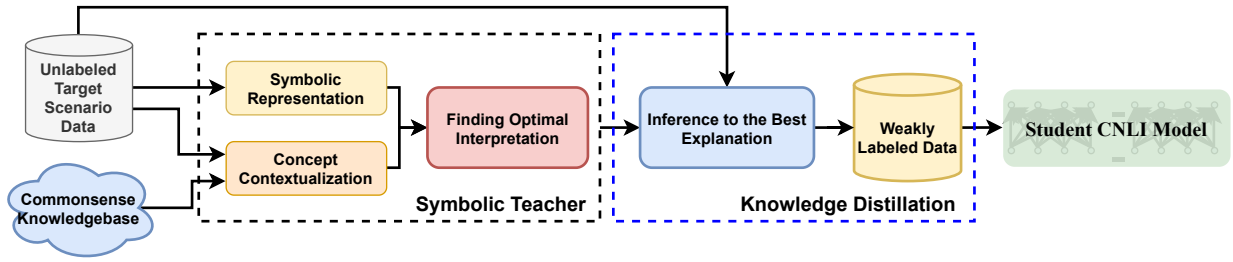

Fig. 2. **Overall approach** is illustrated here. We adopt a hybrid teacher-student framework, with a commonsense knowledge base as the teacher and a CNLI model, trained in a source task, as the student. Given a collection of unlabeled data from the target task, we use a symbolic abductive reasoning framework based on Grenander's pattern theory to create weakly labeled data. A CNLI model is then trained with this weakly labeled data along with an (optional) small labeled data to adapt to a target task.

more, the intermediate graphs generated through pattern theory-based reasoning capture the complex semantic relationships among concepts in each hypothesis to provide an explainable interpretation for understanding its internal reasoning mechanism.

Knowledge distillation was first introduced in [12] and later generalized by [13] as a method to transfer the knowledge learned from larger, more complex models into smaller, more compact networks. This method usually involves training the smaller network (called the student) using soft targets, which are generated by the larger model (the teacher) in addition to the ground truth labels. This allows the soft targets to act as a regularizer and helps to learn better representations. The knowledge distillation framework has been used in various applications such as action recognition [49], visual understanding [50], [51], visual dialog [52], and model compression [53], among others. In the traditional student-teacher framework, the teacher model is usually a large, high-performing model or an ensemble of such models, which is trained in a supervised manner on large-scale training data and used to train smaller, compact student networks. Therefore, the distillation process is more straightforward, where the student is trained on targets provided by the teacher network's predictions. However, in our case, the teacher is a symbolic knowledge base and requires a reasoning mechanism to effectively distill knowledge for a specific task, i.e., CNLI, on unlabeled data. It is to be noted that all these knowledge distillation approaches involve the training of a large teacher network in a *supervised* manner.

Abductive reasoning, introduced by Peirce [11], refers to "*inference to the most plausible explanation for incomplete observations*" and has not been extensively explored in literature from a computational viewpoint. While it is considered to be the source of reasoning used by humans in everyday situations [54], surprisingly few computational models have been introduced. Most of the existing models are logic-based, such as abductive reasoning in formal contexts [55], [56]. A recent abductive reasoning approach is abductive NLI [57], which is framed as supervised question answering.

3 Approach Overview

In this work, we propose a hybrid, unsupervised knowledge distillation approach that uses a symbolic teacher model based on pattern theory to distill general-purpose knowledge from largescale knowledge bases for commonsense natural language inference. In this work, we propose a hybrid, unsupervised approach to distill general-purpose knowledge from large-scale knowledge bases for commonsense natural language inference, using a symbolic teacher model based on pattern theory. In contrast to traditional knowledge distillation applications, the teacher network is not trained in a supervised manner. Instead, we use the idea of abductive reasoning as a mechanism to leverage general-purpose knowledge from symbolic knowledge bases, such as ConceptNet [9], [10], for the CNLI task. The overall approach is illustrated in Figure 2. Given a contextual description E_t and multiple plausible follow-up hypotheses $\{H_n\}$, we formulate an energy-based abductive reasoning framework expressed in Grenander's pattern theory formalism [15] to evaluate the likelihood of each hypothesis and choose the most likely one that completes the observation.

Abductive reasoning typically involves inferring the most plausible hypothesis that completes the observed evidence. This reasoning process typically starts with a set of observations, both complete and incomplete, and attempts to find the most likely explanation for the occurrence of these observations. At the core of this process is commonsense knowledge that evaluates the plausibility of each hypothesis and identifies the hypothesis with the maximum evidence to support its validity. Formally, we define abductive reasoning as an optimization process that aims to find the optimal hypothesis $H_i \in \{H_1, H_2, H_3, \dots H_n\}$ that has the maximum probability of occurrence, conditioned upon the observed evidence E_t and prior commonsense knowledge about the evidence, C_t.This can be expressed as the optimization for

$$\underset{H_i \in \{H_1, H_2, H_3, \dots H_n\}}{\arg\max} \quad p(H_i | C_t, E_t) \tag{1}$$

where E_t represents the observed evidence from the input data at time t. This optimization involves empirically computing the probability of occurrence for each hypothesis H_i given the commonsense knowledge C_t.

In the proposed framework, we represent the context as the observed evidence, the answer candidates as the hypotheses, and ConceptNet as the source of commonsense knowledge. As opposed to logic-based reasoning, we use semantics driven by natural language to drive the reasoning process. Hence, assigning a likelihood for any given hypothesis requires a complete understanding of the observed evidence, which requires interpreting the semantic structure

that links the recognized actors, their actions, and interactions. We express this semantic structure through a graph-based representation called an *interpretation* and express it in terms of Grenander's canonical representation of general pattern theory [15], [58], [59]. Each interpretation is a *contextualized* representation of each hypothesis-evidence pair and conditioned by commonsense knowledge. Evaluating the likelihood of each hypothesis allows us to *weakly* label data from the target task, which can then be used to train CNLI models specific to a given task.

4 Symbolic Teacher: Pattern Theory

At the core of our approach lies the notion of *contextualization*. Contextualization, first defined by Gumperz [60], involves the use of relevant *presuppositions* from prior knowledge to maintain involvement in the current task. More specifically, *presupposition* refers to the inherent knowledge of a concept, such as its properties and shared semantics with other concepts. This allows us to construct interpretations that go beyond simple pairwise relationships and predefined logic and rules. Contextualization has two distinct advantages: (1) it enables us to capture semantic relationships among concepts whose co-occurrence has not been observed, and (2) it helps us move towards an open-world paradigm and bypass the need for annotated training data to learn these semantic associations.

Formally, we represent concepts as g_i for $i = 1, \ldots, N$, and we use ${}_{g_i}R_{g_j}$ to represent semantic relationships between two concepts. Then, the contextualization cue is a concept, g_k, that satisfies the following assertion: not $\left({}_{g_i}R_{g_j}\right) \wedge {}_{g_i}R_{g_k} \wedge {}_{g_k}R_{g_j}$. This means that two concepts that do not have a direct, previously observed relationship can be correlated using contextualization cues. For example, in Figure 4, the use of contextualization cues such as *person*, *music*, and *instrument* allow us to establish a semantic association between the concepts *woman*, *seat*, *nervous*, and *stage*. These interpretations are expressed through a graph-based representation driven by pattern theory [15], [58].

Concepts as Generators: We represent concepts as *generators*, $g_i \in G_s$, where G_s is the collection of all generators required to express the semantics of a given environment. Each generator, g_i, represents a single atomic element that expresses the presence of a concept. We allow for two different types of generators based on their provenance. *Grounded* generators ($\underline{g}_1, \underline{g}_2, \ldots, \underline{g}_q \in G_E$) are concepts whose presence in the interpretation can be grounded to their presence in the evidence. *Ungrounded* generators ($\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_q \in G_C$), on the other hand, represent essential, contextual knowledge about grounded generators. The term *grounding* is used to differentiate concepts based on their presence in the evidence. In Figure 4, the concepts *person, instruments*, and *music* are the ungrounded generators, whereas the other concepts represent the grounded generators. While the ungrounded generators are not directly observed, they are essential to understanding the semantic relationship between the actor (*woman*) and the object of interest (*piano*), moving beyond simple, pairwise semantics.

Expressing Associations using Bonds: Each of the concepts shares a semantic relationship with other generators. These associations can represent specific semantics such as spatial, temporal, and social, to name a few. We express these semantics through links called *bonds*. The direction of the *bonds* signifies the semantics of a concept and the type of relationship shared with its bonded generator. For example, the generators *piano* and *instruments* are semantically related through the assertion that "*a piano is an instrument*". The *energy* of a bond is used to quantify the strength of the semantic relationship expressed between two generators and is given by the function:

$$b_{sem}(\beta'(g_i), \beta''(g_j)) = w_s \tanh(\phi(\beta'(g_i), \beta''(g_j))). \quad (2)$$

where β' and β'' represent the bonds from the generators g_i and g_j, respectively; $\phi(\cdot)$ is the strength of the assertion expressed in the bond; and w_s is a constant used to weight the bond energies. The sentence structure (see Section 4.3), represented by the dependency graph, is used to scale the value of w_s for capturing the structural properties of the sentence, in addition to the semantic properties. We use $\tanh$ to normalize the assertion strength to range from -1 to 1 and hence express both positive and negative assertions. We use ConceptNet [9], [10] as the source of these bonds.

Interpretations as Configurations. The semantics of the observed data are expressed through complex structures called *configurations* (c). Generators combine through their local bond structures. An example of a configuration is shown in Figure 4. Each configuration has an underlying graph topology specified by a connector graph $\sigma \in \Sigma$, where Σ is the set of all available connector graphs. σ, also called the connection type, defines the directed connections between generators. Formally, we define a configuration c as a connector graph σ whose sites $1, \ldots, n$ are populated by generators $g_1, \ldots, g_n$ expressed as,

$$c = \sigma(g_1, \ldots, g_i); g_i \in G_S. \quad (3)$$

The semantic content of the configuration c is defined by the choice of generators $g_1, g_2, \ldots, g_i$. For example, in Figure 4, the sentence "*On stage, a woman takes a seat at the piano. She nervously sets her fingers on the keys.*" can be represented as a configuration (or interpretation) with a set of grounded concepts (stage, woman, nervous, etc.) and ungrounded concepts (person, instrument, and music).

The probability of a given configuration c can be computed by the energy $E(c)$ of the configuration c. The energy of a configuration c is defined as the sum of the bond energies (Equation 2) formed by the bond connections between generators in the configuration and is given by

$$E(c) = -\sum_{(\beta', \beta'') \in c} b_{sem}(\beta'(g_i), \beta''(g_j)) \quad (4)$$

The probability of the configuration is given by $P(c) \propto e^{-E(c)}$. Hence, lower energy indicates higher probability.

4.1 Finding Optimal Interpretations

The large-scale and general nature of commonsense knowledge bases can introduce noise and bias into the reasoning process. Naively considering the energy of the configuration to be the sum of the energies of the semantic bonds can produce very large interpretations, introducing a number of ungrounded generators that are not relevant to the interpretation. To construct interpretations using concepts that

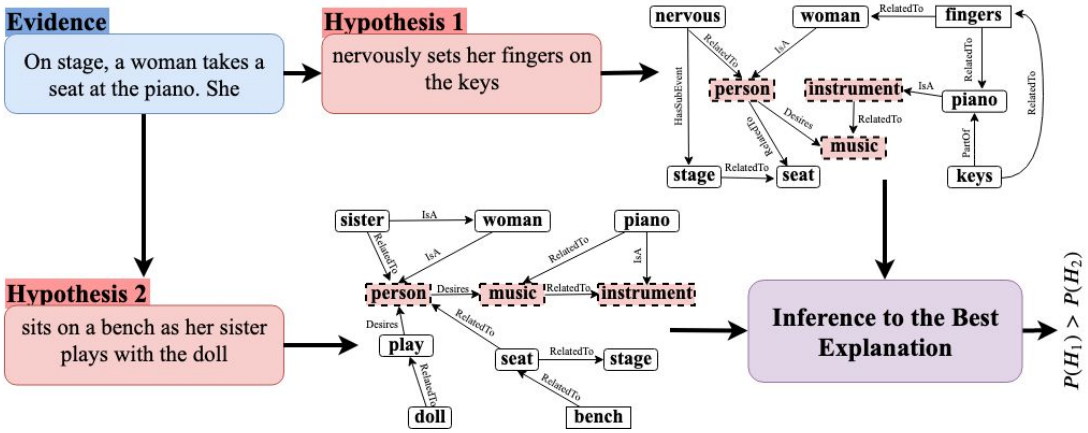

Fig. 3. The proposed **abductive reasoning process** is illustrated here. Given an observed evidence and putative hypotheses, contextualized interpretations are constructed. *Inference to the best explanation* is done using pairwise comparisons to rank the plausibility of the hypotheses.

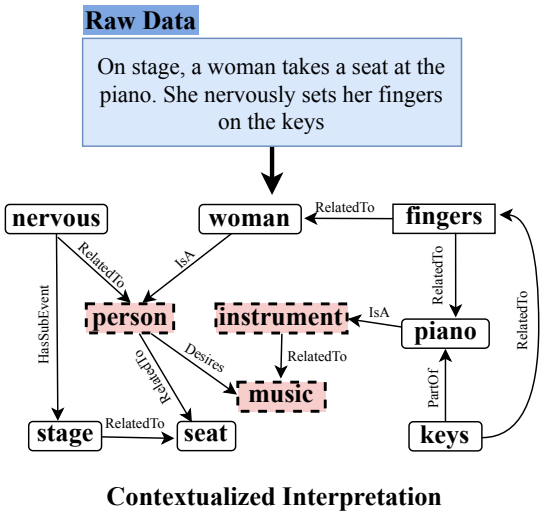

Fig. 4. An example of how natural language sentences are expressed as contextualized interpretations in the pattern theory framework.

are most relevant to the observed evidence, we postulate that the optimal interpretation minimizes the number of ungrounded generators while maximizing its probability.

The process for constructing the optimal contextualized interpretation for a configuration with two grounded generators g_i and g_j is as follows:

1) Extract the subgraph of all related concepts from ConceptNet, representing the contextual properties of the given generators g_i and g_j up to depth d.
2) Construct configurations that represent all grounded concepts and their semantic relationships.
3) Compute the energy of each configuration obtained and find the optimal configuration, i.e., the one with the lowest energy.

The computational complexity of this process is $O(kN^2)$, where k is the number of configurations considered from ConceptNet for each set of N grounded generators. Since we restrict the contextualization to a depth of d, the number of configurations considered is limited. As seen in Table 11, increasing d results in larger configurations, but does not significantly improve the performance.

The task of constructing the contextualized evidence is finding an optimal interpretation, c, given the evidence generators E_t, a set of hypothesis generators H_i, and the prior knowledge in terms of the ConceptNet graph, C_N. We factor this probability into two parts: a likelihood term, $p(G_f|c)$, and a prior, $p(c|C_N)$, normalized by the distribution over the evidence, where $G_f = H_i \bigcup E_t$, the combined set of both evidence and hypothesis generators. The probability of the optimal configuration c can be computed as follows:

$$p(c|C_N, G_f) = \frac{p(G_f|c)p(c|C_N)}{p(G_f|C_N)} \tag{5}$$

This probability can be captured using energy functions:

$$P(c|C_N, G_f) = \frac{1}{Z} e^{-E(G_f|c)-E(c|C_N)} \tag{6}$$

Here, $E(G_f|c)$ represents the energy of the configuration c that involves the grounded generators and the detected concepts, and $E(c|C_N)$ captures the energy of the ungrounded generators. Hence, the total energy $E(c)$ of a configuration c, as defined in Equation 4, is updated to be the sum of these energies: $E(c) = E(G_f|c) + E(c|C_N)$. Each of the terms $E(c|C_N)$ and $E(G_f|c)$ is computed by only summing the energy of all bonds over the ungrounded generators and grounded generators, respectively.

It should be noted that the second term in the exponential $E(c|C_N)$ is not the entire subgraph from ConceptNet but rather the subset that minimizes the overall energy. Hence, the energy of the optimal configuration is given by:

$$E(c|C_N) = \sum_{(\beta', \beta'') \in c} b_{sem}(\beta'(g_i), \beta''(g_j)) + Q(c) \tag{7}$$

where $Q(c)$ is a quality factor that restricts the inference process from constructing configurations with degenerate cases such as unconnected or isolated generators. It is formally defined as

$$Q(c) = k \sum_{\bar{g}_i \in G'} \sum_{\beta^j_{out} \in \bar{g}_i} [D(\beta^j_{out}(\bar{g}_i))] \tag{8}$$

where G' is a collection of ungrounded generators present in the configuration c, β_{out} represents each *out-bond* of generator g_i and D(.) is a function that returns a Boolean value specifying whether the given bond is open i.e., it is not connected to another generator.

We illustrate this process with a simple example. Given a context of a question sentence, "The sun is responsible for," and the answer option (hypothesis) "plants sprouting, blooming, and wilting," we first extract the list of concepts using the NLTK framework and lemmatize them to ensure that we can find them in ConceptNet. We restrict the concepts to nouns, verbs, and adjectives. Hence, G_f is given by sun, responsible, plants, sprout, bloom, wilt. The first step in contextualization is the extraction of the subgraph connecting all grounded concepts and their properties up to a depth d. This results in the extraction of several concepts, some of which are shown in Figure 5. As can be seen, these are all concepts that connect the grounded concepts and can add lots of noise if included as is. The second step is to extract all possible subgraphs that connect all grounded concepts. This can include several possible combinations, some of which are illustrated in Figure 5. We compute the energy of each configuration or subgraph using Equation 5. The third and final step is to find the subgraph with the minimum energy and hence the maximum probability. For this step, we sort the subgraphs by their energies and choose the highest-ranking configuration as the final configuration for the evidence-hypothesis pair. The entire process is shown in Figure 5. It can be seen that it is not a trivial task and hence provides optimal use of the knowledge base for providing a contextualized representation. Note that the configuration on the right has more nodes, and hence a simple sum over the bond energy would result in lower energy. Here, the quality factor restricts the number of ungrounded generators added to the configuration.

4.2 Knowledge Source: ConceptNet

To model the semantics of the interpretations, we use a large commonsense knowledge base as the source of knowledge about concepts and their semantic associations. While our approach is general enough to handle multiple sources of commonsense knowledge [30], [61], [61], we use ConceptNet [9], [10] as the source of **general human knowledge**. ConceptNet is a general-purpose knowledge base that maps concepts and their semantic associations into a large-scale, traversable semantic network. It encodes multi-domain semantic information in a hypergraph, with nodes representing concepts connected through labeled, weighted edges. The semantic relationships between concepts are populated automatically from various sources of knowledge, such as DBPedia [28], Wiktionary, WordNet [29], the OpenCyc ontology [26], and Open Mind Common Sense [62]. ConceptNet contains more than 3 million concepts connected through 34 different assertions (semantic relations), with each assertion specifying and quantifying the semantic relationship between the two concepts, such as `HasProperty`, `IsA`, and `RelatedTo`. Note that the assertion `RelatedTo` expresses a generic, positive semantic relationship between two concepts, while the other named assertions, such as `IsA` and `Has SubEvent`, express specific relationships between concepts. Hence, they may act as a source of noise when using ConceptNet as a source of knowledge. The weight of each edge determines the validity of the assertion. In this work, we consider all the concepts in ConceptNet to be the generator space G_s and quantify the bonds between generators. Hence, the edge weights are used to populate the value of $\phi(\cdot)$ in Equation 2 and determine the validity of the contextualized evidence.

4.3 Capturing Sentence Structure

Creating a contextualized interpretation by simply utilizing the words in the sentences can be too naive and can introduce noise into the reasoning process. To this effect, we use the NLTK framework [63] to parse the sentence and extract the dependency graph between concepts such as nouns and verbs and their associated descriptors such as adjectives and adverbs, respectively. We use the dependency graph to capture the structural associations among these extracted concepts to modulate the semantic relationships as extracted from ConceptNet. We scale the energy of the semantic bond energy, defined in Equation 2, with the dependency structure. The value of w_s is scaled by 0.5 if there is no structural dependency between the concepts and scaled by 1.0 if there is a dependency. Hence, the dependency graph is the initial, underlying graph structure for the interpretation, allowing us to capture the semantics of the question and the answer choice beyond simple, naive semantic relationships between concepts from ConceptNet and reducing the dependency on ConceptNet assertions. Although this seems simple, we see from Section 6 that the use of the dependency graph has a significant impact on the approach's performance when given large sentences that require complex reasoning.

5 CNLI Student: Knowledge Distillation

Our goal is to distill general-purpose knowledge into CNLI models that can function in a given task, given a symbolic teacher framework. We divide this distillation mechanism into two steps. First, we generate *weakly-labeled* training data by utilizing the pattern theory-based abductive reasoning approach detailed in Section 4, using unlabeled data from a target task along with a small amount of labeled data. Second, we use this weakly-labeled data to train a student CNLI model using a knowledge distillation approach. This process helps to reduce the need for supervised training of the teacher network, thus decreasing the requirement for large amounts of labeled data in a target task.

5.1 IBE: Inference to the Best Explanation

The first step in the hybrid knowledge distillation process is the generation of weakly-labeled data. In the abductive reasoning framework, we refer to this step as *inference to the best explanation* since the interpretation with the highest probability is the configuration with the most support from ConceptNet, which is captured in its energy. In our framework, this involves constructing contextualized interpretations for each of the available hypotheses $H_i \in H_n$ along with the observed evidence E_t. The "*plausibility*" of each hypothesis can be obtained by computing the probability of the configuration as defined in Equation 4. Note that

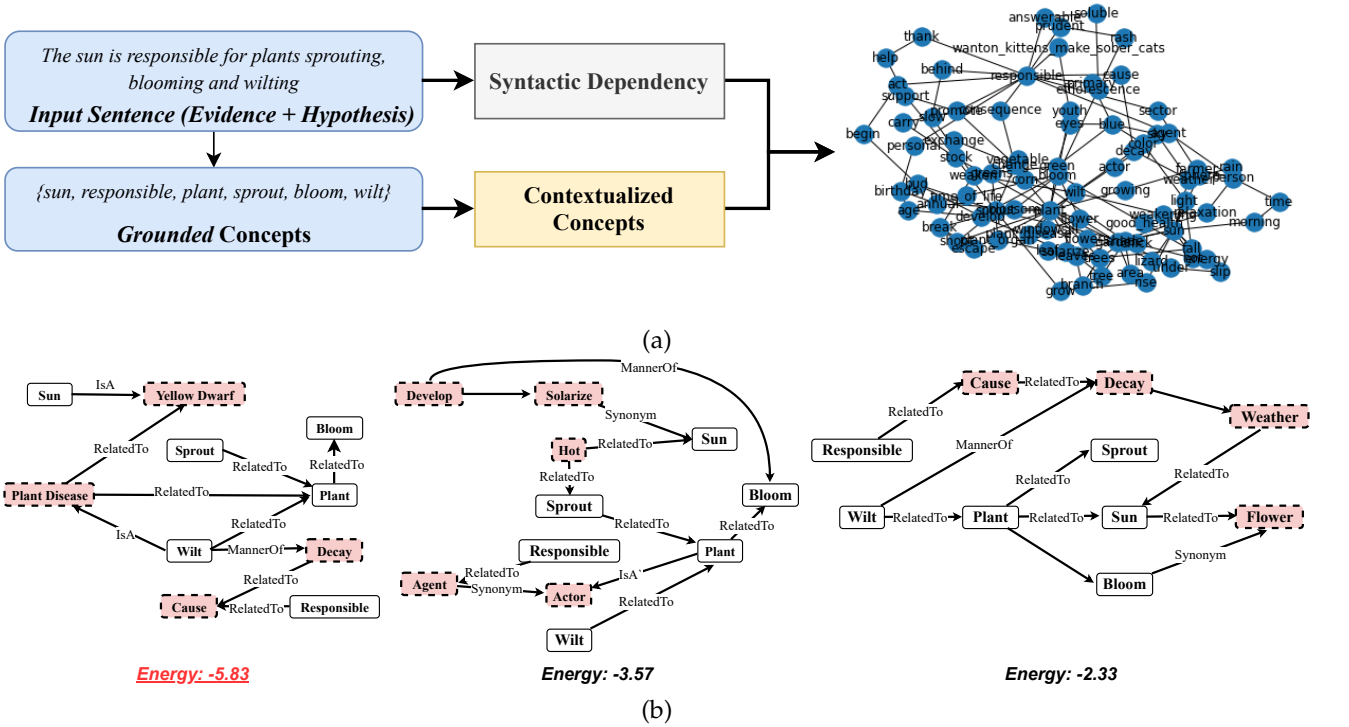

Fig. 5. An illustration of the contextualization process. (a) shows the input evidence and hypotheses and the resulting ConceptNet subgraph that is extracted for reasoning. (b) shows three plausible contextualized interpretations and their corresponding energies. The interpretation with the least energy (first on the left) i.e., highest probability is highlighted in red. Grounded concepts are in white and ungrounded are in red with dotted margins.

a configuration's energy, as defined in Equation 4, is *proportional* to its probability and does not directly provide its probability. To find the *probability* of each configuration, their energies must be normalized using a partition function, which can be intractable since it requires reasoning over all possible configurations that can be present for each hypothesis. Therefore, we use *pairwise comparisons* between the available hypotheses, as illustrated in Figure 3, to find the highest-ranking hypothesis and negate the need for computing the partition function. We use the premise from the Bradley–Terry model [64] to obtain the outcome of the pairwise comparison between two given configurations, as illustrated in Figure 3. The pairwise comparison between configurations c_{H_i} and c_{H_j} is given by Equation 9:

$$P(c_{H_i} > c_{H_j}) = \frac{P(c_{H_i})}{P(c_{H_i}) + P(c_{H_j})} \tag{9}$$

Here, $P(c_{H_i})$ is the probability of the contextualized interpretation of the evidence E_t and a given hypothesis H_i. When this comparison is performed with all available hypotheses H_n, it becomes the optimization for the inference defined in Equation 1. Note that in some instances, there can exist a case of indifference, where two hypotheses can have different configurations with *identical energies*, and hence the probability $P(c_{H_i} > c_{H_j})$ would be 0.5. Any indifference in the outcome is decided by choosing the hypothesis with the highest energy among grounded concept generators. This ensures that the effect of noise introduced through the contextualization process is kept minimal.

5.2 Training Student Models

Our framework allows for training specialist models (BERT, GPT-2, RoBERTa, etc.) for different tasks. However, we would like to point out that the goal of our framework is to distill commonsense knowledge from repositories such as ConceptNet into neural NLI models for faster inference. The pattern theory model (IBE), i.e., the "teacher" model, works unsupervised on all tasks without requiring any training data - synthetic or otherwise [36], [37], [38], and still offers competitive performance to these approaches on all benchmarks. Note that this is not necessarily "zero-shot" since we do not learn a representation or semantic mapping for each domain in order to allow for NLI on different tasks. Given hypotheses and a premise, we ascertain their probability without the need for any kind of training as long as a large-scale, generalized knowledge base such as ConceptNet is present.

We distill the knowledge from the abductive reasoning framework into a specialist neural network (such as BERT or LSTMs) by presenting the hypothesis selected from IBE as the target for optimization. The probability of each hypothesis is given by Equation 10:

$$P(H_i) = \frac{exp(E(c_i)/T)}{\sum_{j}^{n} exp(E(c_j))/T} \tag{10}$$

Here, $E(c_i)$ represents the energy for the given hypothesis H_i, and its corresponding probability is given by $P(c_i)$. T represents the temperature parameter which modulates the probability assigned to each of the target hypotheses. When $T \to \infty$, all hypotheses have uniform probability, and

$T = 1$ represents the standard softmax function. Equation 10 is used to construct the targets for the neural network. We use the energy of each configuration to assign soft-probabilities for the neural network to train on. These soft-probabilities are used in place of one-hot vectors from the ground truth. The temperature function allows us to distill the commonsense knowledge from ConceptNet to the supervised models by enabling us to present cases of indifference from the IBE process (Section 5.1) to the model. This allows us to condition. This allows us to condition the model with cases of semantic indifference, which helps the training process move beyond the structural and co-occurrence-based context.

6 Experimental Evaluation

Data: We evaluate the proposed reasoning approach's performance on three different CNLI datasets spanning various domains. The SWAG [1] dataset consists of 113k multiple-choice questions derived from captions of consecutive events of videos in the ActivityNet Captions [2] and the Large Scale Movie Description Challenge (LSMDC) [3] datasets. The videos cover various domains and hence require reasoning across tasks, temporal scales, and physical interactions to complete the task. The HellaSWAG [5] dataset is another visually grounded CNLI dataset consisting of around 70k multiple-choice questions. It is a more challenging domain introduced by populating question-answer pairs by completing how-to articles from WikiHow. The OpenBookQA [4] dataset is a more challenging CNLI dataset that requires a deeper understanding of both the topic (common sense knowledge) and the language expressed. There are around 6,000 questions based on an "open book" of core, "common sense" facts. We compare two versions of the proposed approach on all datasets. We represent the purely symbolic model as "IBE," whose final label is decided by the reasoning process described in Section 5.1. "PT+BERT" indicates that a BERT model is fine-tuned using the knowledge distillation approach described in Section 5.2, with the labels populated by the "IBE" model. We use the official train, dev, and test split for all datasets.

Challenges: The use of adversarial filtering in SWAG and HellaSWAG datasets ensures that the effect of annotation artifacts is reduced and hence allows us to evaluate the robustness of our approach. These datasets offer three significant challenges: (i) questions go beyond what is observed in natural language and require reasoning across a variety of themes such as physical, social, temporal, and spatial; (ii) language descriptions are grounded in vision, which makes the reasoning over language concepts susceptible to variations in the physical world; and (iii) require a much deeper common sense understanding than simple linguistic entailment for complex, multi-hop reasoning.

6.1 Quantitative Evaluation

We evaluate our approach to help transfer learn CNLI models under different evaluation settings. First, we evaluate the ability of the proposed approach to accelerate the training process of large neural networks like BERT in a *semi-supervised* learning setting, where limited amounts of *labeled* data are available along with large amounts of unlabeled data. Second, we evaluate the performance of the abductive reasoning framework (IBE) in the generalized, zero-shot question-answering setting where no target task data is available for transfer learning. Finally, we evaluate our performance on the *unsupervised* open-domain question answering where the goal is to answer multiple-choice questions *without using domain-specific auxiliary data*.

TABLE 2
Semi-supervised learning results where a limited number of *labeled* data is made available during training. We significantly improve BERT's performance with limited labeled data and unlabeled data.

Labeled Data	OpenBookQA		HellaSWAG	
	BERT	PT + BERT	BERT	PT + BERT
None	27.1	35.6	28.9	30.2
10	26.7	31.6	28.3	28.4
25	26.7	32.8	28.6	28.7
50	26.9	34.0	28.8	29.1
100	27.1	35.8	28.9	29.4
500	28.2	42.2	29.8	30.1
1000	38.8	44.6	30.4	31.1
2500	43.6	45.8	30.8	32.6
ALL	56.6	-	46.7	-

6.1.1 Semi-Supervised Question Answering

We first evaluate the proposed framework for transfer learning under a semi-supervised learning setting, where large amounts of unlabeled data are present along with a small set of labeled data in the target task. We compare it against a fully supervised BERT model that has access to the entire set of labeled data as a baseline. We summarize the results in Table 2. We vary the amount of available labeled data from as few as 10 samples to 2500 samples along with unlabeled examples and evaluate on OpenBookQA and HellaSWAG, two of the more challenging datasets under low data regimes. It can be seen that with as few as 500 labeled samples, we obtain 42.2% accuracy on OpenBookQA, a number which requires $2,500$ labeled samples (50.7% of the training set) for the fully supervised BERT to achieve. Similarly, on HellaSWAG, when 2500 labeled training samples are available, we achieve 32.6% accuracy, which outperforms very strong fully supervised baselines such as a Bidirectional LSTM trained with GloVe embeddings and ESIM with ELMO embeddings (Table 6). Considering that this is only 6% of the training data, this is a remarkable performance and helps significantly reduce the training required for adapting models to novel tasks.

6.1.2 Zero-shot Question Answering

We evaluated the zero-shot ability of language models, such as GPT and GPT-2, and supervised models like BERT by ranking candidate options through computing the likelihood of each option. We calculated the probability of the combined sentence, including both evidence and each hypothesis, and chose the best ranking option as the output. This is a natural baseline for our model, as we replaced the symbolic pattern theory network with the knowledge acquired through the pre-training process. Given that all models were trained on corpora similar to ConceptNet, this

TABLE 3
Evaluation in the **zero-shot setting** on three benchmark data sets

Approach	OpenBookQA	SWAG	HellaSWAG
GPT	26.6	**41.3**	27.8
GPT-2	30.2	40.2	22.9
BERT	27.1	41.1	28.7
IBE	**32.7**	38.4	**28.9**
PT+ BERT	**35.6**	**43.6**	**30.2**

is the closest setting to IBE. We summarized the results in Table 3. IBE, our symbolic reasoning process using ConceptNet as the source of knowledge, outperformed GPT, GPT-2, and BERT in the zero-shot setting by a large margin. It is noteworthy that all three models were trained on corpora similar to ConceptNet and, in BERT's case, trained explicitly for next sentence prediction. Our use of explicit, symbolic representation of commonsense knowledge and contextualized representations allowed us to perform complex reasoning and help generalize to novel tasks without explicit re-training, even when faced with adversarial filtering.

We also evaluated the ability of our approach to train BERT in an unsupervised manner using our PT+BERT approach, where BERT is trained on the task using the knowledge distillation approach in Section 5.2. It can be seen that we consistently improved the ability of BERT to generalize to novel tasks through self-supervised abductive reasoning. We showed that abductive reasoning provided significant gains (9% in absolute accuracy) on OpenBookQA, which required complex and, in some cases, multi-hop reasoning that required a much deeper commonsense understanding than simple linguistic entailment. It is interesting to note that PT+BERT performed better than IBE alone and BERT alone, indicating that the use of knowledge distillation helped capture commonsense assertions beyond pure symbolic reasoning and sequence-based representations.

6.1.3 Unsupervised Transfer Learning for CNLI

Finally, we evaluate our approach on unsupervised CNLI and compare it against baselines with varying degrees of supervision. Our approach does not use any training data; we answer the question by choosing the correct answer choice, purely using ConceptNet as a source of knowledge.

We begin by evaluating on OpenBookQA, which is designed as a benchmark for answering multiple-choice questions about recurring science themes and principles. The dataset is constructed to evaluate the ability to perform question answering using "broad common knowledge," using a set of core facts and an optional set of secondary facts. We compare against four broad types of baselines and summarize the results in Table 4. The first category of baselines consists of systems that rely completely on prior knowledge and use reasoning mechanisms such as self-talk [46], TupleInference [65], and entailment computation (DGEM [66]). We also compare against large, pre-trained language models such as GPT, GPT-2, and BERT to evaluate the use of learned, neural knowledge representation for question answering. Our approach IBE and PT+BERT also belong to this category since we do not use any core facts or additional auxiliary data. In the second category, we compare against models such as KTL [36], MR [37], and consistency optimization [45], which, while not training directly on the data, train auxiliary mechanisms to rewrite questions or use auxiliary, domain-specific prior knowledge for answering questions. In the third category, we allow these approaches to have access to the task-specific set of core facts and the auxiliary data for unsupervised question answering. Finally, we compare against fully supervised models such as ESIM [23], BERT [6], QAGNN [38], OCN [43], and KnowledgePath [39].

It can be seen that we significantly outperform all unsupervised baselines, with and without access to task-specific knowledge, including BERT, GPT, and GPT-2. Our approach (both PT only and PT+BERT) performs competitively with other unsupervised baselines while requiring significantly less overhead for commonsense NLI. For example, KTL [36] requires the construction of a specialist knowledge base geared towards each domain for evaluating each answer option. QASC [40] is used as the source of knowledge for answering questions from OpenBookQA, which is from the same domain as the benchmark and is designed to ensure overlap with the concepts from OpenBookQA. It also requires the translation of questions to hypotheses using a question-specific modifier such as rule-based models to convert wh-questions and answers to statements to evaluate the plausibility of answer choices. Similarly, MR [37] requires the construction of a unified knowledge graph from domains similar to the target domain (they construct a knowledge base called CWWV that utilizes three knowledge bases: ConceptNet, WordNet, and Wikidata) as well as fine-tuning RoBERTa [67] on synthetic QA pairs generated in a sentence in a lexicalization step using a set of pre-defined templates for each type of question along with a distractor sampling step (using RoBERTa embeddings for similarity matching) to prevent the overfitting of the language model to the synthetic QA pairs. Consistency optimization [45] uses various trained mechanisms to translate natural commonsense questions into "fill-in-the-blank" cloze sentences. A language model is then used to compute the probability of each answer choice being the correct answer for the blanks in the sentence. Self-talk [46] performs rewriting of the questions into "clarification" questions conditioned on the context by concatenating pre-defined or generated question prefixes to the context and evaluating the plausibility of each answer choice using a language model.

On the other hand, we do not require the construction of additional specialized knowledge bases, mechanisms for question rewriting, or ensembling for answering questions and can either outperform or provide competitive performance to these approaches. As expected, fully supervised models, augmented with external knowledge, such as QAGNN [38], OCN [43], and KnowledgePath [39], significantly outperform unsupervised and weakly supervised models. Of particular interest is KnowledgePath [39], which generates a path that connects concepts in the question-answer pair from a knowledge graph such as ConceptNet, and each path is scored by GPT-2 [7]. While similar to our approach, its graphs have a chain structure that links each concept to only one other concept and does not capture the semantic dependencies among multiple concepts or move beyond one-hop neighbors as is done in our approach.

TABLE 4
Evaluation on **OpenBookQA**. We outperform unsupervised baselines and offer competitive performance to supervised approaches.

Approach	Dev Accuracy	Test Accuracy
No Training, Only Prior Knowledge		
TupleInference [65]	15.9	17.9
GPT-2 [7]	26.6	-
GPT [68]	30.2	-
BERT [6]	27.1	-
DGEM [66]	27.4	24.4
Self-Talk (GPT-2) [46]	28.4	30.8
IBE (*Ours*)	32.7	-
PT + BERT (*Ours*)	**35.8**	**34.2**
Auxiliary Training, Only Prior Knowledge		
KTL [36]	34.8	34.4
RoBERTa+MR [37]	34.8	38.0
Consistency optimization [45]	**50.3**	**49.9**
No Training, Prior Knowledge + Training Data		
TupleInference [65]	23.6	**26.6**
DGEM [66]	**28.2**	24.6
Fully Supervised Models		
ESIM [23]	53.9	**48.9**
BERT [6]	**56.6**	-
OCN [43]	-	67.8
Knowledge path [39]	-	71.2
QA-GNN [38]	-	**82.8**

TABLE 5
Evaluation on the **SWAG** dataset. We outperform unsupervised, weakly supervised and some early supervised baselines.

Supervision	Approach	Val. Acc.	Test Acc.
None	ConceptNet + Rules [1]	27.0	-
	IBE (*Ours*)	38.4	38.2
	GPT [68]	40.2	-
	GPT-2 [7]	41.3	41.4
	BERT [6]	41.4	-
	PT + BERT (*Ours*)	**43.6**	**43.7**
Weak	DualBoW+GloVe [69]	34.5	34.7
	SNLI + DecompAttn. [69]	35.8	35.8
	SNLI + ESIM [23]	**36.4**	**36.1**
Full	fastText [24]	29.4	28.0
	LSTM + GloVe [1]	43.1	43.6
	ESIM + ELMO [23]	59.1	59.2
	BERT [6]	**86.6**	**86.3**

SWAG. Next, we evaluate our approach on the SWAG dataset, which evaluates the ability of models to perform commonsense natural language inference about visually grounded situations. The dataset is constructed from visually grounded video captions and formulates a CNLI task to predict which event is most likely to occur next in a video. The question is the *context* or the event currently being observed, and the answer choices are the set of plausible events that can follow the current observation. Answering these questions requires general "commonsense" knowledge and an understanding of physical and social dynamics from textual data. Additionally, this data is augmented with adversarial filtering, a mechanism that involves the iterative refinement of hypotheses to present a selection of highly plausible answer choices filtered through counterfactual reasoning. These characteristics pose a challenging benchmark to evaluate our approach to commonsense reasoning.

We compare against a set of baselines with varying levels of supervision. Specifically, we compare against unsupervised baselines such as a simple rule-based reasoning engine using ConceptNet (ConceptNet + Rules) and unsupervised versions of large language models such as GPT, GPT-2, and BERT. We also compare against weakly supervised baselines, which are models trained for textual entailment (i.e., identify entailment, neutral, and contradiction between sentence pairs) on SNLI [69] and fine-tuned for SWAG with these 3-way probabilities as features. Finally, we compare against fully supervised baselines such as fastText [24], ESIM [23], LSTM-based models, and BERT. As shown in Table 5, PT+BERT outperforms all unsupervised baselines by large margins. Interestingly, we also outperform the weakly supervised baselines and early supervised baselines such as fastText and an LSTM-based model with GloVe embeddings. We offer competitive performance to other fully supervised baselines *without any labeled data*.

HellaSWAG. Finally, we evaluate on HellaSWAG, which extends the idea of *grounded* commonsense natural language entailment by presenting answer choices with targeted adversarial filtering. In addition to video captions, HellaSWAG also introduces a new challenge to evaluate commonsense reasoning by framing the CNLI problem to help complete how-to articles from WikiHow, an online how-to manual.

The adversarial filtering is stepped up to a more challenging setting by using GPT-2 as a generating mechanism for alternative answer choices, while BERT is used as a strong discriminator to distinguish between the actual and generated answer choices. The resulting dataset poses a significant challenge for commonsense reasoning that requires both a deep understanding of physical interactions and social situations, in addition to broad commonsense knowledge. We summarize the results in Table 6. We report results for both BERT-base (in italics) and BERT-Large (in parentheses). Note that we only train the Base version in PT+BERT to be consistent with all other approaches. While the language model-based *unsupervised* baselines perform reasonably well on the SWAG dataset, the GPT model performs less than random (22.9%) on the HellaSWAG dataset. GPT-2 achieves 29.5% on HellaSWAG, but considering that the dataset is constructed using the GPT-based model for adversarial filtering, this does not demonstrate the generalization ability of supervised models to newer tasks. We achieve 30.2% on the HellaSWAG dataset with a self-supervised BERT-base model, which is impressive considering that the fully supervised model achieves 39.5%.

This demonstrates the ability to effectively distill knowledge from ConceptNet into neural network models, even with adversarial filtering. In addition to the overall accuracy, HellaSWAG also provides a *zero-shot* setting to evaluate a model's ability to generalize to new situations. The examples in this set are from activity labels from WikiHow and ActivityNet that are *unseen during training*. It is interesting to note that PT+BERT obtains 30.2% on this setting, which is more than fastText (28%) and LSTM+GloVe (29.5%), which are trained under supervised settings, whereas fully supervised BERT-base obtains 36.1%. We perform consistently across all subsets *with no labeled training data* and pose an encouraging way forward to reduce the dependency on labeled data for fine-tuning to a novel task.

TABLE 6
Evaluation on **HellaSWAG**. We outperform all unsupervised baselines, including language models with extensive pre-training.

Supervision	Approach	Val. Acc.	Test Acc.
None	ConceptNet + Rules [1]	20.6	-
	GPT [68]	22.9	25.8
	GPT-2 [7]	27.8	29.5
	BERT [6]	28.7	-
	IBE (*Ours*)	28.9	-
	PT + BERT (*Ours*)	**30.2**	**30.4**
Full	fastText [24]	30.9	31.6
	LSTM + GloVe [5]	31.9	31.7
	ESIM + ELMO [23]	33.6	33.3
	GPT [68]	41.9	41.7
	BERT [6]	*39.5* (**46.7**)	*40.5* (**47.3**)

6.2 Explainability of Pattern Theory Interpretations

In addition to evaluating the performance of the proposed framework, we assess the explainability of the generated interpretations for each question-answer hypothesis. The interpretations offer unique insights into the inner mechanisms of the reasoning process. Since interpretability and explainability are highly subjective, we establish four metrics. Specifically, three objective metrics, *node relevance, edge relevance, and graph completeness*, are used to quantify the relevance of each node and edge to the overall interpretation generated by the pattern theory model, as well as quantifying to the extent which each generated graph provides a complete picture of the question-hypothesis pair. A subjective metric, *overall explainability*, is defined to measure the ability of the generated graphs to express the relationships between the concepts and provides a quantitative metric of the interpretability of the model's internal reasoning mechanism. We describe each metric below.

Node relevance is used to measure the significance of the nodes to understand how all concepts in the sentence are related to each other, including the presence of ungrounded generators. In other words, it assesses the impact of dropping a generator from the interpretation to the semantic coherence of the reasoning graph. *Edge relevance* is defined as a metric to quantify the relevance of the bonds derived from ConceptNet to understand how two concepts are related to each other. In addition, it provides a mechanism to understand how much changing the relationship expressed in a semantic bond can impact the coherence of the interpretation. *Graph completeness* is used to assess the presence of all concepts (i.e., words relating to actions, objects, and their respective qualifiers) in the pattern theory interpretations, in addition to explaining their provenance using potential ungrounded generators. *Overall explainability* is a subjective measure used to quantify the human user's satisfaction with an interpretation's ability to sufficiently capture the underlying semantic structure that connects the hypothesis and the premise.

Evaluation protocol. To assess the explainability of the pattern theory interpretations, we present 100 hypotheses across three datasets, OpenBookQA, SWAG, and HellaSWAG, to 10 human users. Each user is provided with a set of instructions describing the evaluation protocol and a description of metrics. To avoid introducing additional factors such as the accuracy of the answers, we only select hypotheses from the ground-truth question-answer pairs. In addition to graphs generated by our model, we also present graphs from 2 baselines for comparing the explainability of pattern theory interpretations. First, we choose a variation of the proposed approach without contextualization, i.e., considering an interpretation without additional ungrounded generators. These graphs would only consider the concepts in each hypothesis and the direct semantic relationships that are shared by them. Second, we generate a graph with all 2-hop neighbors of concepts from each hypothesis. These graphs are analogous to PT graphs, except they are not optimized to contain only the most relevant ungrounded generators as done in the contextualization process.

TABLE 7
Explainability Studies: We perform user studies to assess the explainability of the proposed approach along with 2 related baselines.

Approach	Node Relevance	Edge Relevance	Graph Completeness	Overall Explainability
IBE	7.45	7.35	7.95	7.85
No contextualization	6.98	7.12	5.86	6.24
2-hop neighbors	5.76	6.85	5.52	4.93

The results of the user study are presented in Table 7. It can be observed that the pattern theory-generated graphs consistently received higher scores from human evaluators than the other two baselines on all metrics, with significantly higher rates when considering the *graph completeness* and *overall explainability* metrics. It should be noted that the approach with no contextualization has comparable *node relevance* and *edge relevance* metrics to the PT graphs since they measure the relevance of the retrieved nodes and edges from ConceptNet to the concepts in the hypothesis but score significantly lower on the *graph completeness* and *overall explainability* metrics which measure the overall significance of the graphs themselves to the interpretability of the graphs. The graphs generated by the 2-hop neighbors' approach introduce many nodes and edges that are not directly relevant to the hypothesis and hence score significantly lower than the other baselines. These results indicate that the contextualization process consistently returns contextually and semantically relevant nodes and edges to the final interpretation that provides greater explainability. It should be noted that while the pattern theory-generated graphs received significantly higher scores than the baselines, there is room for improvement in these metrics for explainability in all approaches. This could arguably be attributed to the fact that there are large amounts of noise and bias from the knowledge bases in the reasoning process.

6.3 Semantic Similarity as Abductive Reasoning

To demonstrate the versatility of the proposed approach, we show that the abductive reasoning framework can be applied to other downstream tasks, such as semantic textual similarity (STS). Semantic textual similarity aims to score the relationship between texts using a defined metric and is a core part of many downstream applications, including information retrieval and text summarization. The most common approach is to learn meaningful representations of sentences in a latent space and use a learned regression model (in

TABLE 8
Semantic similarity: We evaluate the proposed framework on the semantic textual similarity using the STS Benchmark.

Approach	Spearman Rank Coefficient
BERT CLS vector [6]	16.5
Mean-pooled RoBERTa embeddings [67]	12.9
RoBERTa CLS vector [67]	31.7
Mean-pooled BERT embeddings [6]	46.4
IS-BERT-NLI [22]	**69.2**
Ours (IBE)	**42.6**

the case of supervised approaches) or cosine similarity (in the case of unsupervised approaches) to assign a similarity score. Spearman's rank correlation coefficient is used to ascertain the correlation between the similarity scores and human-scored similarity scores from the ground truth. A higher correlation indicates better alignment between human judgment and the model's notion of similarity.

We frame the semantic textual similarity problem as an abductive reasoning task by considering two hypotheses. The default or null hypothesis is that the semantic coherence of the first sentence (the premise) is complete and hence has the lowest energy when contextualized. The addition of concepts degrades the coherence and increases the energy of the configuration. The alternative hypothesis is that the changes in concepts that yield the transformation to the second sentence provide better semantic coherence and hence reduce the energy further, providing reinforcement or entailment for similarity. The resulting energy differential is indicative of the level of similarity between the sentences. The higher the energy differential, the higher the similarity between the two sentences.

We evaluated our approach on the STS Benchmark [17], which consists of sentence pairs labeled from 0 to 5, indicating the level of semantic relatedness. The dataset contains a total of 8628 sentence pairs, with 5749 pairs for training, 1500 for validation, and 1379 for testing. We evaluated directly on the test set without using any training data. For quantitative evaluation, we compared our approach against a variety of recent, unsupervised language model baselines (BERT [6] and RoBERTa [67]) and considered two variations of each language model - representations from the CLS vector and a mean pooled representation from embeddings of each word in the sentence. We also evaluated variations (IS-BERT-NLI [22]) optimized for this task. The resulting embeddings of each sentence were compared using cosine similarity to assign a score for semantic textual similarity. Following prior work [22], we used Spearman's rank correlation between the predicted similarity and the gold labels as the evaluation metric. The results presented in Table 8 show that our approach, although not optimized for this task, outperforms many of the unsupervised baselines and performs competitively with others optimized for this task. The major advantage of our approach is the generation of a contextualized interpretation of the two hypotheses, which offers enhanced explainability (Section 6.2) by providing insight into the model's reasoning process and highlighting potential noise and bias in the knowledge base.

TABLE 9
Sentiment classification: We evaluate the proposed framework with accuracy as a metric on the sentiment classification task using the SST-2 and IMDB Benchmarks.

Approach	Supervision	SST-2	IMDB
BERT [6]	Full	92.3	89.2
MTLE [70]	Full	88.4	91.3
RoBERTa [67]	Full	96.7	95.8
LINDA [71]	5-Shot	63.5	67.3
LINDA [71]	10-Shot	63.5	67.3
DualCL [72]	5-shot	67.1	-
DualCL [72]	5-shot	72.5	-
MTLE [70]	None	71.6	67.5
PT + BERT (Ours)	None	83.5	81.3

6.4 Sentiment Classification

To demonstrate the versatility of the proposed framework beyond NLI tasks, we formulate unsupervised sentiment classification as an abductive reasoning task by considering the labels "positive" and "negative" sentiments as hypotheses (H_i) for a given sentence or phrase, and the evidence (E_t) as input. This setup allows us to adapt our framework for sentiment analysis without significant changes to the overall structure of the hybrid knowledge distillation paradigm. Text classification tasks, such as sentiment analysis, are an integral component of many natural language processing and information extraction frameworks. The prominent approach has been to encode the sentence or phrase using feature extractors such as bag-of-words [24] or embeddings from a language model such as BERT [6]. Then a supervised classifier is trained to make the final prediction about the sentiment of the given sentence. However, few efforts have been made to address this task in an unsupervised manner or under resource-constrained settings.

Several unsupervised techniques have been proposed to tackle sentiment classification. Zhang *et al.* [70] proposed an unsupervised sentiment classification framework using unsupervised matching of learned embeddings to select the most appropriate label for a given sentence. Kim *et al.* [71] proposed LINDA, a data augmentation technique that scales sentiment classification to work in the low training data regime with as few as 5 to 10 labeled examples. Similarly, Chen *et al.* [72] proposed using dual contrastive learning to propose a data augmentation routine for low data sentiment classification. We evaluate our approach on two standard benchmarks, the Stanford Sentiment Treebank (SST-2) [73] and the IMDB Sentiment [18] benchmarks, following prior work. Both datasets evaluate the ability of text classifiers to distinguish between sentences describing positive or negative sentiments sourced from movie reviews.

Table 9 summarizes the performance of our approach and the following comparable baselines. We compare against a variety of fully supervised, weakly (few-shot) supervised and unsupervised baselines and report the average accuracy as the quantitative performance metric. Specifically, we user a fully-supervised BERT [6] and RoBERTa [67] as the fully supervised large language model baselines, as well as the weakly supervised models such as LINDA [71] and DualCL [72]. We compare against both the unsupervised and fully supervised versions of MTLE [70]. As can

be seen from Table 9, we outperform both unsupervised and weakly supervised baselines while offering competitive performance to the fully supervised approaches. Interestingly, we achieve 83.5% accuracy on SST-2, while a fully supervised BERT achieves 92.3%. This performance is in line with the performance of the hybrid knowledge distillation approach on other tasks and datasets, where we obtain more than 75% of the performance of a fully supervised BERT without using any labeled training examples. The pattern theory framework is able to effectively leverage the knowledge from symbolic knowledgebases such as ConceptNet to provide supervision for unsupervised sentiment classification, even with the limited context provided by the single-word labels.

6.5 Zero-shot Text Classification

As a final litmus test, we evaluate the generalization capabilities of the proposed abductive reasoning framework to tackle the problem of *zero-shot* text classification, which is a core part of many NLP and information extraction frameworks. Zero-shot text classification aims to correctly assign a pre-defined yet unseen label to a given span of text. Large language models such as GPT-2 [7] and GPT-3 [8], as well as masked language models such as BERT [6] and RoBERTa [67], have provided powerful baselines for this task due to their ability to capture contextual information in their word embeddings, which is gleaned from pre-training on large amounts of text corpora. The common approach to zero-shot and few-shot learning using these models is through "*prompting*", a method to transform any task, such as text classification, into a language modeling or masked language modeling problem. This method works by inserting pre-defined (both learned and manually assigned) "templates" of text for *prompting* the language model to complete the sentence to provide the required classification task. The other form of zero-shot transfer to new tasks is the idea of *in-context* learning (ICL) [8], where a short description of the task, along with a set of examples, is presented to the model for few-shot adaptation. These are natural baselines to compare against our approach, which works by contextualizing (analogous to "prompting") a symbolic knowledge base (i.e., ConceptNet) for addressing the problem of text classification. Note that we do not claim to perform prompting on symbolic knowledge bases exactly like large language models, but instead, provide a proof-of-concept example of how the abductive reasoning framework can be adapted to a novel task. We leave the problem of tackling general-purpose neuro-symbolic "prompting" to future work since it is beyond the scope of the current work.

We evaluate our approach on two standard benchmarks: RTE [74], [75], [76], [77] and TREC-6 [78]. RTE is a dataset proposed as a standard benchmark for generic semantic inference required in many essential tasks, such as information retrieval, question answering, and information extraction. Framed as a text classification task, the goal is to identify whether the meaning of one sentence can be inferred from another. TREC-6 is a multi-class, text classification dataset consisting of open-domain question-answer pairs that need to be classified as belonging to one of six coarsely labeled classes. Performance on both datasets is quantified with accuracy. We compare against zero-shot and few-shot versions of GPT-2 [7] and GPT-3 [8], as evaluated by Zhao *et al.* [79]. We also compare against the different variations of zero-shot learning using RoBERTa [67], as reported by Gao *et al.* [80]. For a fair comparison, we only compare against the vanilla versions of prompting and in-context learning, which use a learned language model in place of a symbolic knowledge base, as is the case with our approach.

TABLE 10
Zero-shot Text classification: Generalization ability is evaluated on the text classification tasks, as evaluated on the RTE and TREC-6.

Approach	Supervision	TREC-6	RTE	Avg
GPT-2	0-Shot	24.0	51.0	37.5
	1-Shot	21.5	57.6	39.6
	4-Shot	23.1	53.2	38.2
	8-Shot	32.7	54.9	43.8
GPT-3	0-Shot	31.0	44.8	37.9
	1-Shot	24.3	49.6	36.9
	4-Shot	25.8	44.0	34.9
	8-Shot	29.3	49.2	39.3
RoBERTa (Prompting)	0-Shot	32.0	51.3	41.7
RoBERTa (ICL)	0-Shot	26.2	60.2	43.3
RoBERTa	Full	97.4	80.9	89.2
IBE (Ours)	0-Shot	31.6	53.8	42.7
PT + BERT (Ours)	None	**57.3**	**62.4**	**59.9**

As shown in Table 10, the proposed abductive reasoning framework, referred to as *IBE*, performs well on the zero-shot setting where there is no fine-tuning on the target dataset domain. We outperform most zero-shot and few-shot baselines, with only the zero-shot version of RoBERTa using prompting outperforming our approach. It is interesting to note that we outperform all few-shot baselines except GPT-2 in the 8-shot setting. When an unlabeled dataset is available for training, the proposed hybrid knowledge distillation approach outperforms all few-shot baselines while achieving an average accuracy of 59.9% across the two tasks. Remarkably, this is 67.1% of the performance of a fully supervised RoBERTa model. Although there is a relatively large gap between the performances of the supervised and unsupervised approaches, it is encouraging to see that the proposed approach provides a significant first step in closing the gap by leveraging large-scale knowledge bases without any labeled data.

6.6 Ablative Studies

In addition to quantitative analysis, we systematically evaluate the different components of the proposed approach. Specifically, we evaluate three specific components: (i) the effect of contextualization, (ii) the source of semantic knowledge, and (iii) the student or specialist model. Table 11 summarizes the results of the ablation study.

Effect of Contextualization. First, we evaluate the impact of the use of contextualization (Section 4.1) on the overall performance of the proposed approach. We use different variations of the contextualization approach by varying the context depth d from 0 (i.e., without contextualization) to 5, which indicates that we look for semantic assertions between two concepts up to the depth $d = 5$. As shown in Table 11, when d=0, the performance drops drastically to

TABLE 11
Ablative Studies: We compare different source of knowledge and different student networks. Evaluation results are reported on SWAG.

Approach	Val. Accuracy
Effect of Contextualization	
No Contextualization (d=0)	33.6
Context depth d=1	34.8
Context depth d=2	35.3
Context depth d=3	37.2
Context depth d=4 (Full Model)	38.4
Context depth d=5	38.9
Different Knowledge Sources	
Numberbatch Only	25.9
GloVe Only	26.3
GloVe + Numberbatch	28.1
IBE (No Sentence Structure)	35.2
Different Student Models	
PT + LSTM+GloVe	32.4
PT + ESIM+GloVe	39.4
PT + BERT	**43.7**

33.6%, which is a gap of 6.3%. Each increment in the context depth d yields improvements, with the best performance at d=5. After depth d=4, the inference time increases non-linearly and does not yield significant improvements in accuracy. Hence, our final model uses a depth of $d = 4$, which provides a balance between inference time and accuracy. The use of contextualization to construct interpretations yields improvements of 6.3% in accuracy.

Here's the cleaned up paragraph:

Source of Semantics. Our framework can handle different sources of knowledge, but we primarily use ConceptNet's symbolic knowledge and NLTK's syntactic knowledge (Section 4.3). To evaluate the performance with other knowledge sources, we vary the source of semantic knowledge by using GloVe [81] representations and ConceptNet NumberBatch [10]. The strength of the assertion ($\phi(\cdot)$ from Equation 2) is computed using the dot-product between the vector embedding of the two concepts, which allows us to evaluate the use of contextual word embedding instead of symbolic knowledge for unsupervised QA in the Pattern Theory framework. Table 11 shows that ConceptNet, along with contextualization, is essential for robust commonsense reasoning. ConceptNet Numberbatch, which is trained on ConceptNet, does not provide the same performance as ConceptNet as a symbolic knowledge base. Using representations learned from pre-computed embeddings such as GloVe or Numberbatch without ConceptNet assertions does not generalize to the QA task. The use of the semantic dependency graph (Section 4.3) to capture the sentence structure also yields significant gains (3.2%) and shows that pattern theory representations can integrate multiple sources of knowledge into the reasoning process without manual curation of rules for reasoning.

Different Student Models. Besides BERT, we train two student networks: ESIM and a Unary LSTM model. The LSTM baseline takes an arbitrary span of text (question + answer choice) as input and encodes it using a two-layer Bidirectional LSTM network. The hidden state of the LSTM network is then max-pooled to obtain a fixed-size representation, which is used to obtain a probability of occurrence for that answer choice. The ESIM model is pre-trained on SNLI with ELMo embedding. The output entailment prediction layer is replaced with a new classification layer to predict the probability of co-occurrence of the question and the specified answer choice. Table 11 shows that BERT achieves the highest accuracy, but the LSTM model with GloVe embedding obtains 32.4% accuracy when trained in an unsupervised manner with the predictions from IBE and knowledge distillation. Compared to the fully supervised performance of 43.1%, the performance of the LSTM student model is remarkable and represents 75% of the supervised model's performance. Similarly, ESIM trained with ELMo embedding obtains 39.4% accuracy, compared to 59.1% from the fully supervised version. These results show that our framework can be used to train a variety of student models and still perform competitively with fully supervised baselines.

7 Limitations and Future Work

While the approach performs well on different CNLI tasks (Section 6), as well as on other downstream tasks such as semantic similarity (Section 6.3) and sentiment classification (Section 6.4), we observe that the framework has some limitations and specific error modes that can be the focus of future work to improve the abductive reasoning mechanism. For example, we note that the performance gap between the fully supervised model and our approach reduces as the complexity of the model decreases. The knowledge distillation approach (Section 5.2) as well as the inherent noise from the weak-labeling in the pattern theory framework (Section 5.1) add a measure of regularization. However, we still observe that the addition of *labeled* data does not always result in increased performance. This effect was acute in the semi-supervised learning setting (Table 2), where it took more than 100 labeled training examples, in addition to the unlabeled data, to outperform the completely unsupervised transfer using PT+QA. This effect could arguably be attributed to the fact that larger models such as BERT tend to pick up on spurious patterns in the data and tend to overfit certain training examples [1], [5]. Further regularization techniques [82] can help mitigate this effect.

The other key limitation of the approach is the possible propagation of noise and bias from the knowledge bases into the reasoning process. ConceptNet is a large, general-purpose knowledge base that spans various domains. It captures *concept*-based semantic relationships mined from a wide variety of sources. Hence, there is a strong potential for the injection of noise into the reasoning process, particularly by generic assertions such as *RelatedTo*, which do not provide specific, verified semantic relationships between concepts. We limit this effect by defining a strong constraint using the contextualization process, where the additional context depth improves the accuracy of the underlying pattern theory reasoning framework. However, we find that noise seeps into the process, as indicated by the relatively lower explainability scores (Table 7), although it does outperform comparable baselines. Some examples are

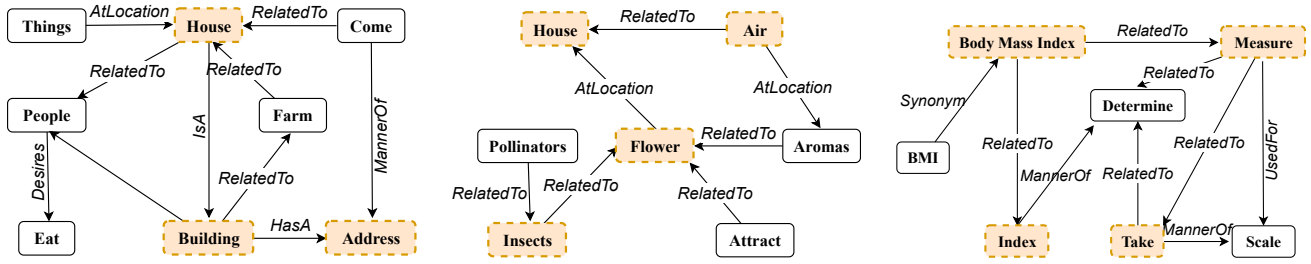

Fig. 6. **Qualitative Examples** of the generated interpretations that highlight the impact of noise that is inherent in large-scale knowledge bases such as ConceptNet that can impact the contextualization process. Ungrounded generators are shaded and the predicted answer is underlined.

shown in Figure 6. For example, in the example on the right, while the contextualization process correctly equated *BMI* with "body mass index", there are some unnecessary concepts such as *index* that add noise to the interpretation. This is much more acute in the other middle example, where the concepts "house" and "flower" were forced into the interpretation while not directly related to the query. Some ungrounded generators introduced due to noise or bias in the knowledge base can greatly affect the framework's performance, particularly on those with adversarial filtering, such as HellaSWAG. Other mechanisms, such as affordance constraints [83], can help further mitigate this effect. Similarly, the contextualization process has additional computational overhead since it requires reasoning over possible subgraphs connecting the grounded concepts from ConceptNet. Using graph generative transformers [84], [85] can help reduce the computational overhead by learning to sample contextualized subgraphs from ConceptNet.

Finally, our approach is designed for tasks where the hypotheses are predefined and the goal is to select the correct hypothesis. Extensive experiments have demonstrated that the approach can be used for various tasks that follow this general problem setup. However, its potential applications to generative tasks such as translation or summarization have not been explored in this work. We envision its use in grounding and constraining the outputs of generative models to enhance their semantic coherence, factual correctness, and interpretability. Our future work aims to expand the scope of the abductive reasoning process to include multimodal grounding and event comprehension beyond text-based semantics, moving towards open-world reasoning with limited training requirements.

8 Conclusion

In this work, we present one of the first attempts to distill symbolic knowledge from large-scale knowledge bases for task transfer in commonsense natural language inference. Based on the notion of abductive reasoning and hybrid knowledge distillation, we show that a global source of commonsense knowledge can be distilled into neural networks without requiring large amounts of annotations. We demonstrate the use of pattern theory to express the evidence in a highly interpretable and contextualized interpretation for validating the plausibility of natural language expressions, without training highly expensive models. Extensive experiments demonstrate the applicability of the approach to different tasks, such as commonsense natural language inference (CNLI), sentiment classification, text classification, and semantic textual similarity, and its highly competitive performance with respect to fully supervised transfer learning baselines. We aim to extend the framework for general-purpose neuro-symbolic reasoning over multimodal data.

Acknowledgments

This research was supported in part by the US National Science Foundation grants CNS 1513126, IIS 1956050, and IIS 1955230. The final draft paper was edited for grammar and language using Grammarly and ChatGPT.

References

[1] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 93–104.

[2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 706–715.

[3] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision (IJCV)*, vol. 123, no. 1, pp. 94–120, 2017.

[4] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2381–2391.

[5] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4791–4800.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[9] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.

[10] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *The AAAI Conference on Artificial Intelligence*, 2017.

[11] C. S. Peirce, *Collected papers of Charles Sanders Peirce*. Harvard University Press, 1931.

[12] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 535–541.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[14] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "Comet: Commonsense transformers for knowledge graph construction," in *Association for Computational Linguistics (ACL)*, 2019.

[15] U. Grenander, *Elements of pattern theory*. JHU Press, 1996.

[16] U. Grenander, M. I. Miller *et al.*, *Pattern theory: from representation to inference*. Oxford University Press, 2007.

[17] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," *arXiv preprint arXiv:1708.00055*, 2017.

[18] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language technologies*, 2011, pp. 142–150.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392.

[20] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa: Fact-based visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 10, pp. 2413–2427, 2017.

[21] G. Bekoulis, C. Papagiannopoulou, and N. Deligiannis, "A review on fact extraction and verification," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–35, 2021.

[22] Y. Zhang, R. He, Z. Liu, K. H. Lim, and L. Bing, "An unsupervised sentence embedding method by mutual information maximization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1601–1610.

[23] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced lstm for natural language inference," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1657–1668.

[24] A. Joulin, É. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 427–431.

[25] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2249–2255.

[26] D. B. Lenat, M. Prakash, and M. Shepherd, "Cyc: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks," *AI Magazine*, vol. 6, no. 4, pp. 65–65, 1985.

[27] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*, 1998, pp. 86–90.

[28] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web*. Springer, 2007, pp. 722–735.

[29] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[30] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3027–3035.

[31] X. Li, A. Taheri, L. Tu, and K. Gimpel, "Commonsense knowledge base completion," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1445–1455.

[32] J. Davison, J. Feldman, and A. M. Rush, "Commonsense knowledge mining from pretrained models," 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)*, 2019, pp. 1173–1178.

[33] B. D. Trisedya, J. Qi, W. Wang, and R. Zhang, "Gcp: Graph encoder with content-planning for sentence generation from knowledge base," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[34] L. Lin, Y. Gao, K. Gong, M. Wang, and X. Liang, "Graphonomy: Universal image parsing via graph reasoning and transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[35] Y. Li, B. Cui, and Z. M. Zhang, "Efficient relational sentence ordering network," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[36] P. Banerjee and C. Baral, "Self-supervised knowledge triplet learning for zero-shot question answering," in *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*. Association for Computational Linguistics (ACL), 2020, pp. 151–162.

[37] K. Ma, F. Ilievski, J. Francis, Y. Bisk, E. Nyberg, and A. Oltramari, "Knowledge-driven data construction for zero-shot evaluation in commonsense question answering," in *35th AAAI Conference on Artificial Intelligence*, 2021.

[38] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "Qa-gnn: Reasoning with language models and knowledge graphs for question answering," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 535–546.

[39] P. Wang, N. Peng, F. Ilievski, P. Szekely, and X. Ren, "Connecting the dots: A knowledgeable path generator for commonsense question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4129–4140.

[40] T. Khot, P. Clark, M. Guerquin, P. Jansen, and A. Sabharwal, "Qasc: A dataset for question answering via sentence composition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8082–8090.

[41] B. Y. Lin, X. Chen, J. Chen, and X. Ren, "Kagnet: Knowledge-aware graph networks for commonsense reasoning," 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)*, 2019, pp. 2829–2839.

[42] Y. Feng, X. Chen, B. Y. Lin, P. Wang, J. Yan, and X. Ren, "Scalable multi-hop relational reasoning for knowledge-aware question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1295–1309.

[43] K. Ma, J. Francis, Q. Lu, E. Nyberg, and A. Oltramari, "Towards generalizable neuro-symbolic systems for commonsense question answering," in *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 2019, pp. 22–32.

[44] Y. Xu, C. Zhu, S. Wang, S. Sun, H. Cheng, X. Liu, J. Gao, P. He, M. Zeng, and X. Huang, "Human parity on commonsenseqa: Augmenting self-attention with external attention," *arXiv preprint arXiv:2112.03254*, 2021.

[45] Z.-Y. Dou and N. Peng, "Zero-shot commonsense question answering with cloze translation and consistency optimization," *arXiv preprint arXiv:2201.00136*, 2022.

[46] V. Shwartz, P. West, R. Le Bras, C. Bhagavatula, and Y. Choi, "Unsupervised commonsense question answering with self-talk," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4615–4629.

[47] A. Tamborrino, N. Pellicanò, B. Pannier, P. Voitot, and L. Naudin, "Pre-training is (almost) all you need: An application to commonsense reasoning," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3878–3887.

[48] P. Ramamurthy and S. N. Aakur, "Isd-qa: Iterative distillation of commonsense knowledge from general language models for unsupervised question answering," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1229–1235.

[49] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 12, pp. 2799–2813, 2017.

[50] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[51] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[52] D. Guo, H. Wang, and M. Wang, "Context-aware graph inference with knowledge distillation for visual dialog," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[53] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in

Advances in Neural Information Processing Systems, 2017, pp. 742–751.
[54] H. R. Fischer, "Abductive reasoning as a way of worldmaking," *Foundations of Science*, vol. 6, no. 4, pp. 361–383, 2001.
[55] C. Elsenbroich, O. Kutz, and U. Sattler, "A case for abductive reasoning over ontologies." in *OWLED*, vol. 216, 2006.
[56] J. Meheus and D. Batens, "A formal logic for abductive reasoning," *Logic Journal of IGPL*, vol. 14, no. 2, pp. 221–236, 2006.
[57] C. Bhagavatula, R. Le Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W.-t. Yih, and Y. Choi, "Abductive commonsense reasoning," in *International Conference on Learning Representations (ICLR)*, 2019.
[58] S. Aakur, F. de Souza, and S. Sarkar, "Generating open world descriptions of video using common sense knowledge in a pattern theory framework," *Quarterly of Applied Mathematics*, vol. 77, no. 2, pp. 323–356, 2019.
[59] S. N. Aakur, F. D. M. de Souza, and S. Sarkar, "Going deeper with semantics: Exploiting semantic contextualization for interpretation of human activity in videos," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.
[60] J. J. Gumperz, "Contextualization and understanding," *Rethinking context: Language as an interactive phenomenon*, vol. 11, pp. 229–252, 1992.
[61] J. L. Martinez-Rodriguez, I. Lopez-Arevalo, and A. B. Rios-Alvarado, "Openie-based approach for knowledge graph construction from text," *Expert Systems with Applications*, vol. 113, pp. 339–355, 2018.
[62] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2002, pp. 1223–1237.
[63] E. Loper and S. Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.
[64] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
[65] T. Khot, A. Sabharwal, and P. Clark, "Answering complex questions using open information extraction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 311–316.
[66] ——, "Scitail: A textual entailment dataset from science question answering," in *The AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[67] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[68] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.
[69] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 632–642.
[70] H. Zhang, L. Xiao, W. Chen, Y. Wang, and Y. Jin, "Multi-task label embedding for text classification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4545–4553.
[71] Y. Kim, S. Jeong, and K. Cho, "Linda: Unsupervised learning to interpolate in natural language processing," *arXiv preprint arXiv:2112.13969*, 2021.
[72] Q. Chen, R. Zhang, Y. Zheng, and Y. Mao, "Dual contrastive learning: Text classification via label-aware data augmentation," *arXiv preprint arXiv:2201.08702*, 2022.
[73] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.
[74] I. Dagan, O. Glickman, and B. Magnini, "The pascal recognising textual entailment challenge," in *Machine Learning Challenges Workshop*. Springer, 2005, pp. 177–190.
[75] R. B. Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor, "The second pascal recognising textual entailment challenge," in *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, vol. 7, 2006.
[76] L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo, "The fifth pascal recognizing textual entailment challenge." in *TAC*, 2009.
[77] D. Giampiccolo, B. Magnini, I. Dagan, and W. B. Dolan, "The third pascal recognizing textual entailment challenge," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007, pp. 1–9.
[78] E. M. Voorhees and D. M. Tice, "Building a question answering test collection," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, pp. 200–207.
[79] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 697–12 706.
[80] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 3816–3830.
[81] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
[82] W. Xu, X. Cheng, K. Chen, and T. Wang, "Symmetric regularization based bert for pair-wise semantic reasoning," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1901–1904.
[83] S. N. Aakur, S. Kundu, and N. Gunti, "Knowledge guided learning: Open world egocentric action recognition with zero supervision," *Pattern Recognition Letters*, vol. 156, pp. 38–45, 2022.
[84] D. Belli and T. Kipf, "Image-conditioned graph generation for road network extraction," *arXiv preprint arXiv:1910.14388*, 2019.
[85] S. Kundu and S. N. Aakur, "Iterative scene graph generation with generative transformers," *arXiv preprint arXiv:2211.16636*, 2022.

Sathyanarayanan N. Aakur received the B.Eng. degree in Electronics and Communication Engineering from Anna University, Chennai, India in 2013. He received the M.S. degree in Management Information Systems and the Ph.D. degree in Computer Science from the University of South Florida, Tampa, in 2015 and 2019, respectively. He is currently an Assistant Professor with the Department of Computer Science at Oklahoma State University since 2019. His research interests include commonsense reasoning for visual understanding, multimodal event understanding and deep learning applications for genomics. He received the US National Science Foundation CAREER award in 2022. He has been an Associate Editor with the IEEE Robotics and Automation Letters since 2021.

Sudeep Sarkar is a Distinguished University Professor, Chair of Computer Science and Engineering at the University of South Florida, Tampa, and Co-Director of the USF Institute for Artificial Intelligence + X. He received his M.S. and Ph.D. in electrical engineering on a University Presidential Fellowship from The Ohio State University, Columbus, and his B. Tech degree from the Indian Institute of Technology, Kanpur. He has 35 years of experience conducting and directing fundamental research in computer vision, predictive learning, gait biometrics, and artificial intelligence. He has directed 22 Doctoral and 25 Master's students on these topics. He is a co-Editor-in-Chief of Pattern Recognition Letters and was the President of the IEEE Biometrics Council. He is Fellows of IEEE, IAPR, AAAS, AIMBE, and IAPR and member of the Academy of Science, Engineering, and Medicine of Florida.