ISD-QA: Iterative Distillation of Commonsense Knowledge from General Language Models for Unsupervised Question Answering

Priyadharsini Ramamurthy, Sathyanarayanan N. Aakur*
Department of Computer Science
Oklahoma State University
Stillwater, OK 74078
{pramamu,saakurn}@okstate.edu

***Abstract*—Commonsense question answering has primarily been tackled through supervised transfer learning, where a language model pre-trained on large amounts of data is used as the starting point. While successful, the approach requires large amounts of labeled question-answer pairs, with increasingly larger amounts of data required as the complexity of scenarios or tasks such as commonsense QA increases. In this paper, we hypothesize that large-scale pre-training of language models encodes the necessary commonsense knowledge to answer common questions in context without labeled data. We propose a novel framework called Iterative Self Distillation for QA (ISD-QA), which extracts the "dark knowledge" encoded during largescale pre-training of language models to provide supervision for commonsense question answering. We show that the approach can be used to train common neural QA models for commonsense question answering by distilling knowledge from language models in an unsupervised manner. With no bells and whistles, we achieve an average of 68% of the performance of fully supervised QA models while requiring no labeled training data. Extensive experiments on three public benchmarks (OpenBookQA, HellaSWAG, and CommonsenseQA) show the effectiveness of the proposed approach.**

I. Introduction

Language models, driven by increasingly bigger transformer-based [1] architectures and large amounts of pre-training data, have demonstrated incredible progress in natural language generation [2], [3], relation extraction [4], visual question answering [5], [6], [7], and dialog generation [8], to name a few. However, such success has largely been driven by supervised transfer learning, where large amounts of *labeled* training data are carefully curated to perform these tasks successfully. However, as the task becomes increasingly complex, such as answering questions about common, everyday facts grounded in commonsense knowledge and physical interactions, the amount of data required for such models can be very large. This dependency on labeled training data restricts the ability of such models to truly demonstrate an understanding of commonsense knowledge beyond rote memorization.

Different from other question answering tasks [9], [10], [11], commonsense question answering requires a broader understanding of commonsense knowledge to go beyond semantic entailment and domain-specific knowledge. There is a need for reasoning over physical, semantic, and even social relationships between concepts found in common, everyday events, and facts. While there have been efforts to leverage external knowledge sources such as ConceptNet [12], the focus has still been placed on supervised or semi-supervised learning approaches to commonsense question answering. Additionally, complex reasoning and careful knowledge curation are required for these approaches to help imbue commonsense into language understanding models. There have been few works [13], [14], [15] that have addressed the need for unsupervised, general commonsense question answering, which is an essential part of true machine intelligence since it is unreasonable to assume the availability of large amounts of labeled, curated training data for all possible scenarios.

In this work, we aim to answer the question: "*Can pre-trained language models learn to answer commonsense questions with minimal external support?*" To this end, we hypothesize that language models such as GPT [16], GPT-2 [2] and even BERT [17] implicitly capture these commonsense relationships in their internal framework that can be used for general commonsense question answering. Specifically, we argue that, in addition to their contextualized word representations, the pre-training setup, such as objective functions, inherently encode commonsense relationships between words beyond semantic coherence. Building on the idea of expectation-maximization [18] and knowledge distillation [19], [20], we propose to extract this "*dark knowledge*" inherently captured in pre-trained language models to learn to answer commonsense questions. We demonstrate that language models encode some commonsense knowledge in their pre-training mechanisms, including their objective functions, which can be exploited to provide self-supervision for learning to answer commonsense questions. We evaluate our approach on three different benchmark datasets (CommonsenseQA [21], OpenBookQA [22] and HellaSWAG [23]) that each represents three different scenarios (or contexts) of commonsense knowledge such as semantic relationships between common concepts, elementary science facts, and everyday events and how-to instructions. Combined, these three datasets allow us to evaluate the different facets of

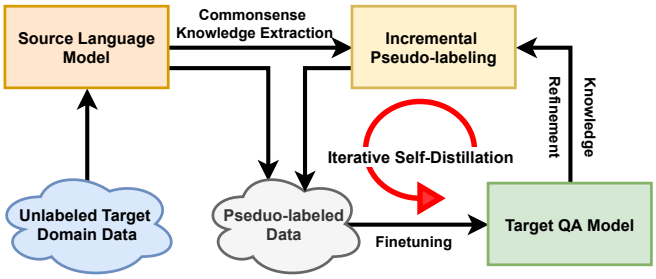


Fig. 1. An **overview** of the proposed ISD-QA architecture is illustrated. Given unlabeled data, we extract and iteratively refine commonsense knowledge from language models for unsupervised commonsense QA.

commonsense knowledge that can be embedded in language models during pre-training.

The **contributions** of our work are three-fold: (i) we are among the first to demonstrate that pre-trained language models can be used to provide supervision for general commonsense question answering without *labeled* training data, (ii) introduce a novel learning mechanism called Iterative Self-Distillation (ISD-QA) to distill commonsense knowledge from language models into QA models, and (iii) show that the proposed ISD-QA approach can work under unsupervised and semi-supervised settings to train QA models with only a language model and minimal external support.

II. Related Work

Commonsense question answering has largely been tackled through supervised transfer learning where a pre-trained language model [17], [24], [25], [26] is finetuned on labeled data in a specific commonsense scenario. Several benchmark datasets have been facilitated for such tasks such as OpenBookQA [22], CommonsenseQA [21], HellaSWAG [23], and aNLI [27], to name a few. Efforts have shifted to incorporating external knowledge from large, general-purpose knowledgebases [28] for commonsense question answering, but most have focused on supervised learning. Efforts have recently focused on unsupervised commonsense question answering by leveraging knowledgebases [29] and language models for cloze translation-based QA [13], [14], [15]. Although these approaches do not need labeled data for question answering, they do need supervision for the auxiliary task of Cloze translation as well as complex reasoning mechanisms. We, on the other hand, focus on extracting commonsense knowledge from language models with *minimal external input* through iterative refinement of their task and vocabulary.

Knowledge distillation (KD) is a methodology proposed by Bucilua *et al.* [20] as a mechanism for effective model compression and expanded by Hinton *et al.* [19] to transfer the knowledge embedded in one neural network to another without losing its validity. Using a student-teacher mechanism, KD has generally been used for distilling a large teacher network into a smaller student network that performs the same task as the teacher. Common applications include language modeling [30], [24], information retrieval [31], visual understanding [32], [33], visual dialog [34], quantization [35], spoken question answering [36] and model compression [37], [38] to name a few. Aakur *et al.* proposed a hybrid knowledge distillation approach for distilling knowledge from symbolic knowledgebases such as ConceptNet [12] for commonsense question answering. On the other hand, we propose to distill general-purpose commonsense knowledge from pre-trained language models using an iterative approach based on expectation-maximization without any external reasoning mechanisms or knowledge sources to incrementally refine its performance on unsupervised commonsense question answering.

III. ISD-QA for Unsupervised Commonsense QA

In this section, we describe the proposed framework, Iterative Self-Distillation for Question Answering (ISD-QA), for unsupervised commonsense question answering with generalized language models. We begin by defining the problem statement and motivation. We then introduce an energy-based framework for commonsense question answering by leveraging the pre-training tasks that are used for training general-purpose language models. We then describe the iterative self-distillation approach for fine-tuning the language model in a transfer learning setting for commonsense question answering. Finally, we describe the implementation details for completeness and reproducible research.

Problem Statement. We consider the problem of unsupervised commonsense question answering, where there exists an unlabeled dataset $X{=}\{x_1, x_2, \ldots x_N\}$ in a given target scenario, where each example x_i is a combination of a question (or context) x_i^q and a set of plausible answers $x_i^A{=}\{x_i^{a_1}, \ldots x_i^{a_j}\}$ to the question. The goal, then, is to leverage the commonsense knowledge embedded in general language models (GLM) [16], [2], [3], [17] to learn to answer the questions with no human supervision. The overall process is illustrated in Figure 1. We consider a language model to be any function that learns the likelihood of occurrence of the words that can occur in a specific sentence. Typical language models have been based on neural networks such as long short-term memory networks (LSTMs) [39], and more recently based on transformers [1]. The language models are typically pre-trained on large corpora of natural language in a self-supervised manner using objective functions ($\psi(x_i, \theta)$, where θ refers to the parameters of the GLM) that maximize the log-likelihood of words occurring in valid, observed sentences. This is usually achieved by many strategies, common among which are next word prediction, masked language modeling [17], next sentence prediction [17] or a combination of them. We propose to use these pre-training objective functions ($\Psi{=}\psi_1(x_i, \theta) \ldots \psi_k(x_i, \theta)$) to iteratively distill commonsense knowledge for question answering using two steps: (i) specialized QA and (ii) iterative self-distillation. We explain each step of the process in detail below.

A. Specialized QA from Generalized Language Models

The first step in our approach is to devise a mechanism to provide a probability of a possible answer choice $x_i^{a_j}$

being correct for a specific question x_i^q using the commonsense knowledge embedded in a language model $\mathcal{M}$, its parameters θ and any $k >= 1$ pre-training objective functions $\Psi=\{\psi_1, \psi_2, \dots \psi_k\}$. Given a question-answer pair $x_i=(x_i^q, x_i^{a_j})$, we compute its probability of being correct by

$$p(x_i^{q,a_j}|x_i^q, x_i^{a_j}, \mathcal{M}) = \sum_{m=1}^{k} \frac{w_m \rho_m(x_i^q, x_i^{a_j}, \theta)}{\sum_{m=1}^{k} w_m} \quad (1)$$

where ρ_m is a function that transforms the output of the objective function ψ_m into a probability distribution and w_m is a weight associated with each objective function ψ_m. In our experiments, we set $w_m = 1$ for all objective functions that can be present in a language model. We set $\rho_m(\cdot)$ to be linear if the objective function ψ_m is binary cross-entropy, as is the case with the next sentence prediction task in BERT and an exponential function if it is the log-likelihood in case of next word prediction (as in GPT and GPT-2) or masked language modeling (as in BERT). The output probability from Equation 1 provides a measure of the confidence with which the language model perceives the given statement (question+answer) to be a valid distribution based on the prior knowledge embedded during pre-training. The final answer for a given question x_i^q and a set of possible answers $x_i^A=\{x_i^{a_1}, \dots x_i^{a_j}\}$ is decided through pairwise comparison between the different answer choices. For example, given two answer choices $x_i^{a_1}$ and $x_i^{a_2}$, we compute $P(x_i^{a_1}>x_i^{a_2})=\frac{p(x_i^{q,a_1})}{p(x_i^{q,a_1})+p(x_i^{q,a_2})}$, based on the probability function defined in Equation 1. Note that in some instances, there can exist a case of indifference, i.e., two answer choices can have *equal probabilities*. These cases of indifference are decided by choosing the answer with the highest probability when stop words and out of vocabulary words are removed. This ensures that any noise due to unknown word sequences is reduced. Given this mechanism, we create an initial training dataset with the pseudo-labels generated for each question-answer choice pair using the unlabeled training dataset in the target task or domain.

B. Iterative Self-Distillation for Transfer Learning

The second step in our approach is to transfer learn the target QA model to the target domain using iterative self-distillation. The goal of this step is to adapt the QA model to the target domain's semantics and vocabulary for specialized QA. We combine the ideas of knowledge distillation and Expectation-Maximization algorithm [18] for iterative self-distillation. The algorithm has four components: (i) initialization, (ii) incremental pseudo-labeling, (iii) self-distillation, and (iv) exit condition. We describe each step in detail below.

Step 1: Initialization: We first initialize the parameters for the target QA model in the target domain by training the QA model for question answering using the pseudo-labeled data produced using the QA specialization described in Section III-A. These steps provide an initial estimate for the QA model's parameters, denoted by θ_0, and allow us to begin the transfer learning process for question-answering in the target domain without human-annotated data. Next, we train the QA model for one epoch with a learning rate of λ_0 and parameters initialized with those of the original language model $\mathcal{M}$. However, it can be any target QA model.

Step 2: Incremental pseudo-labeling. Next, we fix the parameters of the QA model (θ_{t-1}) from the previous step and generate new pseudo-labels for each question-answer pair in the original unlabeled training dataset. We use the original probabilities assigned by the source language model $\mathcal{M}$ using Equation 1 and combine it with the new probabilities generated by the QA model at time t with parameters θ_{t-1} using a weighted average. This step allows us to incrementally update the pseudo-labels with the new, contextualized word embeddings that are finetuned to the target domain semantics.

Step 3: Self-distillation. Given the pseudo-labels generated in the previous step, we then continue to finetune the QA model with the updated, labeled training set. This step allows us to further update the parameters θ_t and the underlying model $\mathcal{M}_{t-1}$ to ensure that they capture the necessary context present in the new scenarios that may not have been present during the pre-training stage. Note that this step is analogous to the maximization step in EM, where we estimate the updated parameters to maximize the likelihood function. At each time step, the learning rate λ_t is updated such that $\lambda_t=0.5*\lambda_{t-1}$ to ensure that the model parameters are not updated too rapidly.

Step 4: Exit condition. Steps 2 and 3 are iterated consecutively during every epoch of the transfer learning process to finetune the QA model to the target domain until the likelihood is not converged i.e., the pseudo-labels between steps $t-1$ and t continue to change. Once converged, the iterative self-distillation process is complete, and the training is stopped.

Combined, these four steps help extract commonsense knowledge from a generalized language model and train a QA model for commonsense question answering. This iterative process is designed to converge to stable, local minima on the training set (through step 2) and prevent overfitting of the target QA model through the self-distillation process (step 3). A visualization of this is shown in Figure 2(a), where the training accuracy converges to a stable state where the labels are no longer changing, and the validation accuracy also increases before converging. Empirically, we show that the QA models trained using this iterative process can generalize better to the test set (Section IV-B). Additionally, we can see that without the iterative pseudo-labeling, the accuracy of the QA model can be significantly lower ($\tilde{6}\%$) when using the QA model with only the initialization step. The iterative process does help the model adapt to target domain semantics and tasks without human supervision in the form of labeled data. Note that the choice of the language model and QA model is arbitrary, although we keep them the same for our primary model. In Section IV-C, we show that the approach can work for any choice of source and target model.

C. Implementation Details

We use the Huggingface Transformers library [40] for our experiments. We experiment with three different language

models, GPT, GPT-2, and BERT as the source of commonsense knowledge and train BERT and RoBERTa as the target QA model. Our final model uses BERT as both the source language and the target QA models. The initial learning rate was set to 4e-4 and reduced by half at every epoch. All language models are used as pre-trained by their original authors and the pre-trained models are used as the initial QA model. The experiments were conducted on a server with an AMD Epyc processor with 32 cores and 2 NVIDIA Titan RTX GPUs. A batch size of 16 (on each GPU) and gradient accumulation of 8 were used to have an effective batch size of 512. The iterative distillation was set to continue until convergence or a maximum of 10 epochs. Hyperparameters were kept constant for all experiments.

IV. Experimental Evaluation

In this section, we describe the evaluation setup, such as datasets used and comparative baselines and evaluation outcomes of the proposed approach on various setups including those with challenges such as adversarial filtering and unseen vocabulary. We also perform ablative studies to systematically examine the various components of the proposed approach.

A. Evaluation Setup

Data. We evaluate our approach on three benchmark datasets, each covering a different domain or scenario in commonsense question answering - CommonsenseQA [21], OpenBookQA [22] and HellaSWAG [23]. The statistics for each of the datasets are shown in Table I. The datasets cover a wide variety of commonsense scenarios, including knowledge about concept-concept relationships, elementary science facts, everyday events, and common how-to questions. Combined, these datasets offer a comprehensive evaluation of the general commonsense knowledge embedded in language models during pre-training. We use the official train, dev, and test split for all datasets. A short description of each dataset and its challenges are presented below.

TABLE I
Dataset statistics for commonsense question answering

Dataset	Context Scenario	Train	Dev	Test	Choices
CommonsenseQA	Concept Relationships	9,741	1,221	1,140	5
OpenBookQA	Science Facts	4,957	500	500	4
HellaSWAG	Events and How-to	39,905	10,042	10,002	4

The CommonsenseQA [21] and OpenBookQA [22] are commonsense question-answering datasets designed to evaluate QA systems with questions that require the use of commonsense knowledge and a reasoning mechanism to answer without contexts, such as semantic relationships between common concepts and elementary science facts. Each question is filtered and presented with distractor answer options that require broad commonsense knowledge and a deeper understanding of concepts beyond simple entailment for complex, multi-hop reasoning. The HellaSWAG [23] dataset is a multiple-choice, commonsense question-answering dataset, grounded in vision by constructing the question-answer pairs from two sources: captions of videos from ActivityNet [41] and by completing how-to articles from WikiHow, an online how-to manual. The HellaSWAG [23] dataset is designed to evaluate the commonsense knowledge embedded in QA systems to reason and predict events that are most likely to occur next, given an observation of the current event. Furthermore, the dataset offers an additional challenge in the form of adversarial filtering, which ensures that the effect of annotation artifacts is reduced and allows us to evaluate the robustness of our approach.

TABLE II
Evaluation on **CommonsenseQA** with comparisons against supervised and unsupervised baselines

Approach	Supervision	Test Accuracy
self-talk (ALBERT)	✗	17.49
LM1B-Concat	✗	25.3
LM1B-Rep	✗	26.1
self-talk (GPT2)	✗	**31.53**
QACOMPARE+GloVe	✓	25.7
ESIM+NumberBatch	✓	30.1
QA-BILINEAR+GloVe	✓	31.5
ESIM+GloVe	✓	**32.8**
BIDAF++	✓	32.0
GPT	✓	45.5
BERT	✓	**52.6**
ISD-BERT (Ours)	✗	**33.1**

Evaluation Metrics and Baselines. We follow the official split and report accuracy for all datasets. We compare against a variety of supervised, weakly supervised, and unsupervised baselines to evaluate the performance of our approach. We compare with several unsupervised baselines such as LM1B-Concat [22], LM1B-Rep [22], and self-talk [13], which forms some of the direct comparisons to our work, since they also propose to leverage language models pre-trained on large corpora for unsupervised QA. We use self-talk with GPT2 [2] and ALBERT [25] as representative baselines for self-talk. We also compare against PT+BERT [29], which propose to train BERT using symbolic knowledgebases such as ConceptNet [12] as supervision using an energy-based Pattern Theory framework. We also compare with fully supervised, transformer-based baselines such as BERT [17], GPT [16], GPT2 [2], and ALBERT [25], as well as traditional baselines such as ESIM [42], QA-BILINEAR [22], QA-COMPARE [22], and BIDAF++ [22], [43] to demonstrate the gap between unsupervised and supervised approaches to QA. Note that we report accuracy only for some baselines on the validation set, since access to the test set is restricted through limited submissions to the leaderboard.

B. Quantitative Evaluation

We evaluate our approach primarily on commonsense question answering tasks presented in CommonsenseQA and OpenBookQA. Both datasets assess the ability of QA models to answer questions about common, everyday facts about semantic relationships and basic scientific facts. These knowledge

TABLE III
EVALUATION ON **OPENBOOKQA** WITH COMPARISONS AGAINST BASELINES WITH VARYING LEVELS OF SUPERVISION.

Approach	Supervision	Dev Accuracy	Test Accuracy
self-talk (ALBERT)	None	22.2	19.4
PMI	None	19.7	21.2
PT + BERT	None	35.8	34.2
self-talk (GPT2)	None	28.4	30.8
TupleInference	Weak	23.6	**26.6**
DGEM	Weak	**28.2**	24.6
ESIM	Full	53.9	**48.9**
BERT	Full	**56.6**	-
ISD-BERT (Ours)	None	**39.8**	**36.0**

elements are not explicitly encoded within general language models (GLMs), and hence these datasets offer an ideal testbed for evaluating their ability to encode commonsense knowledge. Table II and Table III summarize the performance of the proposed approach on CommonsenseQA and OpenBookQA, respectively. It can be seen that we generally outperform all unsupervised, weakly supervised, and some early supervised baselines by a significant margin on both datasets. It is a remarkable performance considering that the source of knowledge for answering the questions is not explicitly provided for comparable unsupervised QA models, such as PMI [22], DGEM [22], IR [22] and TupleInference [22] on OpenBookQA and supervised approaches, such as QACOMPARE, ESIM, and QA-BILINEAR on CommonsenseQA. These results indicate that large general language models encode significant amounts of commonsense knowledge during the pre-training stage with their self-supervised objective functions that can effectively answer questions from previously unseen scenarios. Of particular interest is the comparison to other unsupervised learning approaches that leverage language models as the source of knowledge - self-talk [13] and LM1B. Self-talk uses language models to answer questions by iteratively querying language models to discover any additional background knowledge given the question as context, in the form of a Cloze translation task. LM1B uses a language model [21] trained on the One Billion Words Benchmark [44] to answer questions by using the LM to rank options as either concatenating each question-answer pair (LM1B-Concat) or using heuristics to initialize the ranking through semantic similarity (LM1B-REP). We significantly outperform all baselines on both datasets while using no external, handcrafted systems to extract the commonsense knowledge from GLMs other than the objective function used during pre-training.

We also compare against fully supervised baselines and outperform earlier approaches such as QACOMPARE, QA-BILINEAR, and ESIM, and offer competitive performance to modern, transformer-based approaches. Specifically, we can achieve 33.1% accuracy on CommonsenseQA *without any labeled data* using a BERT-Base model. In contrast, a fully supervised model achieves 52.6%, which indicates that we can achieve 63% of the performance of the fully supervised BERT model. Similarly, on the OpenBookQA dataset, our unsupervised baseline can obtain 36%, while the fully supervised version achieves 56.6%, which is 64% of the fully supervised model's performance. This performance is remarkable considering that BERT-like models are shown to be rapid surface learners [23], that require large amounts of *labeled* training data to adapt to a given domain, particularly in commonsense QA. It is interesting to note that it requires the availability of more than 50% of the labeled training data for a supervised model to outperform the proposed ISD-QA approach. This is illustrated in Figure 2(b) and Figure 2(c) for CommonsenseQA and OpenBookQA, respectively. Similarly, when a fraction of the training data is labeled, our approach outperforms the supervised baselines significantly while ultimately converging to a similar performance when all labeled training data is available (a performance difference of 0.5% on average).

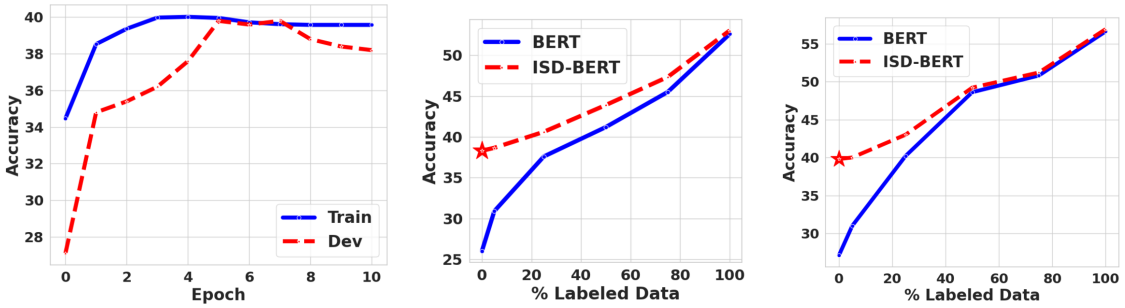
TABLE IV
EVALUATION ON **HELLASWAG** WITH ADVERSARIAL FILTERING

Supervision	Approach	Dev Accuracy	Test Accuracy
None	ConceptNet + Rules	20.6	-
	PT+BERT	30.2	30.4
	ISD-BERT (Ours)	**31.3**	**31.8**
Full	FastText	30.9	31.6
	LSTM+GloVe	31.9	31.7
	ESIM+ELMO	33.6	33.3
	BERT	**39.5**	**40.5**

We also investigate the effect of Adversarial Filtering (AF) and longer sentences on the proposed approach by evaluating on HellaSWAG, which contains answers with more than $5\times$ the number of words in OpenBookQA and CommonsenseQA. Adversarial Filtering (AF) is an interesting data collection paradigm employed in HellaSWAG to generate plausible answers to questions focused on common, everyday events. The question-answer pairs are generated by training a series of strong discriminator QA models in an adversarial manner against a generator language model, generating increasingly plausible wrong answers. In HellaSWAG, the answers are generated with GPT-2 as the generator model and BERT as the discriminator model by scaling up the length and complexity of examples where the text is explicitly made strong against language models. The results are summarized in Table IV, where it can be seen that we outperform other unsupervised baselines and early supervised baselines such as FastText and LSTM+GloVe while offering competitive performance to ESIM. Interestingly, we achieve 31.8% accuracy in an unsupervised manner, which is 78.5% of the performance of a fully supervised BERT model (40.5%). This performance is remarkable, considering that the answer options are generated adversarially, designed to confuse BERT. The approach achieves 31.3% on the zero-shot sections of the test set, which is unseen during the pre-training and transfer learning stages.

C. Ablative Studies

Our primary model, ISD-BERT, uses BERT as both the source language model and target QA model since it provides



(a) (b) (c)

Fig. 2. Visualization of (a) Training-vs-Dev accuracy through Iterative Distillation on OpenBookQA, (b) Performance of a semi-supervised ISD-BERT compared to supervised BERT on CommonsenseQA, (c) Performance of a semi-supervised ISD-BERT compared to supervised BERT on OpenBookQA

TABLE V
ABLATION STUDIES ON DIFFERENT SOURCE LMS AND TARGET QA MODELS. RESULTS ON OPENBOOKQA.

Source LM	Target QA Model	Dev Accuracy
GPT	BERT	35.4
	RoBERTa	36.2
GPT-2	BERT	37.6
	RoBERTa	37.1
BERT	BERT	**39.8**
	RoBERTa	39.6

an elegant, self-training mechanism. However, our approach is general enough to accommodate any source LM and target QA models. To this end, we systematically examine the important components of the proposed approach, such as the source language model and the target QA model, through ablation studies on OpenBookQA. Specifically, we explore language models trained with different objective functions such as left-to-right masking (GPT and GPT-2) and masked language modeling with next sentence prediction (BERT) and other target QA models such as RoBERTa and BERT. As it can be seen from Table V, the performance of the different combinations remain remarkably consistent, showing that the iterative training process is stable and can be generalized to various settings for commonsense question answering. This is further highlighted in Figure 2(a), where it can be seen that the training accuracy converges after a few iterations, while the dev (validation) accuracy quickly catches up by transfer learning the underlying language model to the target scenario's semantics. It can be seen that the training accuracy starts to stagnate after a few epochs of iterative distillation, while the initial accuracy already outperforms other unsupervised baselines. The later epochs, i.e., when the accuracy stagnates, act as a regularizer for the target QA model and help it generalize to the test set. This effect can be seen in Table III and Table IV, where the difference between dev and test accuracy is minimal (3%) compared to supervised approaches (6%), and on the zero-shot section of HellaSWAG, where ISD-BERT obtains 31.3% compared to the overall performance of 31.8%. This demonstrates that the proposed iterative self-distillation approach allows for generalized commonsense question answering using only a language model's initial pre-training objective functions with limited human supervision.

V. DISCUSSION AND FUTURE WORK

In this work, we presented ISD-QA, an unsupervised question answering framework that leverages the commonsense knowledge embedded in language models pre-trained on large corpora. We demonstrate that general language models capture commonsense knowledge beyond just their embedding. The pre-training objective functions also provide a powerful mechanism to extract this knowledge for commonsense question answering. We observe that the iterative distillation mechanism for generating pseudo-labels and finetuning QA models acts as a regularization mechanism that helps prevent overfitting and can help train models under unsupervised and semi-supervised settings. Extensive experiments on three public benchmark datasets show that the approach is effective to help train QA models for commonsense question answering with zero human-annotated labeled data and provides a way forward for generalized commonsense question answering. The trained QA model outperforms other unsupervised, weakly supervised, and some early supervised baselines significantly and offers competitive performance to transformer-based fully supervised baselines. On average, ISD-QA can help train BERT to achieve 68% of the performance of a fully supervised BERT model across three different benchmark datasets spanning diverse commonsense scenarios and challenges such as Adversarial Filtering. We aim to leverage these results to train multimodal commonsense reasoning models to leverage external knowledge for open-world multimodal question answering.

ACKNOWLEDGMENT

This research was supported in part by the US National Science Foundation grants IIS 1955230 and IIS 2143150.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[3] 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.

[4] T. A. N. Dang, D. T. Hoang, Q. B. Tran, C.-W. Pan, and T. D. Nguyen, "End-to-end hierarchical relation extraction for generic form understanding," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5238–5245.

[5] M. Farazi, S. Khan, and N. Barnes, "Question-agnostic attention for visual question answering," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3542–3549.

[6] A. U. Khan, A. Mazaheri, N. D. V. Lobo, and M. Shah, "Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering," *arXiv preprint arXiv:2010.14095*, 2020.

[7] N. Messina, F. Falchi, A. Esuli, and G. Amato, "Transformer reasoning network for image- text matching and retrieval," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5222–5229.

[8] P. Budzianowski and I. Vulic, "Hello, it's gpt-2-how can i help you? towards the use of pretrained language models for task-oriented dialogue systems," *EMNLP-IJCNLP 2019*, p. 15, 2019.

[9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *EMNLP*, 2016.

[10] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *ACL*, 2017.

[11] Y. Yang, W. tau Yih, and C. Meek, "Wikiqa: A challenge dataset for open-domain question answering," in *EMNLP*, 2015.

[12] H. Liu and P. Singh, "Conceptnet — a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, pp. 211–226, 2004.

[13] 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.

[14] 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.

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

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[18] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

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

[20] 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.

[21] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "Commonsenseqa: A question answering challenge targeting commonsense knowledge," in *Proceedings of NAACL-HLT*, 2019, pp. 4149–4158.

[22] 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 *EMNLP*, 2018.

[23] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *ACL*, 2019.

[24] V. SANH, L. DEBUT, J. CHAUMOND, and T. WOLF, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[25] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2019.

[26] 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.

[27] 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*, 2019.

[28] A. Mitra, P. Banerjee, K. K. Pal, S. Mishra, and C. Baral, "Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering," *ArXiv*, vol. abs/1909.08855, 2019.

[29] S. N. Aakur and S. Sarkar, "Abductive reasoning as self-supervision for common sense question answering," *arXiv preprint arXiv:1909.03099*, 2019.

[30] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," *ArXiv*, vol. abs/1909.10351, 2020.

[31] G. Izacard and E. Grave, "Distilling knowledge from reader to retriever for question answering," *arXiv e-prints*, pp. arXiv–2012, 2020.

[32] 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.

[33] 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.

[34] 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.

[35] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," in *International Conference on Learning Representations (ICLR)*, 2018.

[36] C. You, N. Chen, and Y. Zou, "Knowledge distillation for improved accuracy in spoken question answering," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7793–7797, 2021.

[37] 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.

[38] Z. Yang, L. Shou, M. Gong, W. Lin, and D. Jiang, "Model compression with two-stage multi-teacher knowledge distillation for web question answering system," *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[41] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[42] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced lstm for natural language inference," *arXiv preprint arXiv:1609.06038*, 2016.

[43] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[44] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," *arXiv preprint arXiv:1312.3005*, 2013.