Transferring Domain-Agnostic Knowledge in Video Question Answering

Tianran Wu[1]
tianran.wu@is.ids.osaka-u.ac.jp

Noa Garcia[1]
noagarcia@ids.osaka-u.ac.jp

Mayu Otani[2]
otani_mayu@cyberagentco.jp

Chenhui Chu[3]
chu@i.kyoto-u.ac.jp

Yuta Nakashima[1]
n_yuta@ids.osaka-u.ac.jp

Haruo Takemura[1]
takemura@cmc.osaka-u.ac.jp

[1] Osaka University
Osaka, Japan

[2] CyberAgent, Inc.
Japan

[3] Kyoto University
Kyoto, Japan

Abstract

Video question answering (VideoQA) is designed to answer a given question based on a relevant video clip. The current available large-scale datasets have made it possible to formulate VideoQA as the joint understanding of visual and language information. However, this training procedure is costly and still less competent with human performance. In this paper, we investigate a transfer learning method by the introduction of domain-agnostic knowledge and domain-specific knowledge. First, we develop a novel transfer learning framework, which finetunes the pre-trained model by applying domain-agnostic knowledge as the medium. Second, we construct a new VideoQA dataset with $21,412$ human-generated question-answer samples for comparable transfer of knowledge. Our experiments show that: (i) domain-agnostic knowledge is transferable and (ii) our proposed transfer learning framework can boost VideoQA performance effectively.

1 Introduction

Video question answering (VideoQA), which is designed to predict an answer to a given question based on a relevant video clip, requires comprehending both visual and linguistic content. This has made VideoQA an ideal testbed to evaluate current machine learning models, and researchers have made great efforts to advance the field. For example, TVQA [15] presents a large-scale dataset and a model that leverages Faster-RCNN [21] and LSTMs [12] to process visual and language inputs, and the use of attention mechanisms [24] has also achieved a great success [30, 34, 35]. Recently, a new research direction in VideoQA has emerged, *i.e.*, external knowledge-based VideoQA [7, 8], which requires information that

cannot be directly obtained from the videos or the question-answer (QA) pairs, and thus, cannot be learned from the dataset. In this task, therefore, a model needs to refer to knowledge from external sources. This configuration may be closer to real-world question answering.

Knowledge for VideoQA, including external knowledge-based one, can be obtained in different ways: some knowledge can be learned from the videos and QA pairs in the dataset, and other knowledge can be extracted from a (external) knowledge base. Existing approaches use a dedicated set of sentences, each of which is associated with its respective QA pair [8] or a block of text found in the Internet [7, 16], as knowledge base in a specific domain, which are retrieved by the model during inference. This means that the knowledge base can be potentially replaced without re-training. In this case, a natural question arises here: *How much is a trained VideoQA model's knowledge generalised for different domains?*

In this paper, we study the transferability of knowledge learned in VideoQA. Specifically, we split the knowledge into two categories (Figure 1): *domain-specific* and *domain-agnostic*. The domain-specific knowledge is related to a particular domain of videos in a VideoQA dataset. This type of knowledge is acquired for each domain via training. On the other hand, the domain-agnostic knowledge involves (i) the so-called tacit knowledge, such as how to understand videos and questions as well as how to retrieve the desired piece of knowledge from the knowledge base and (ii) the so-called explicit knowledge that is still valid for different domains, such as common-sense.

Based on this distinction, we propose a knowledge-oriented transfer learning for better generalisation of VideoQA models. We study the influence of knowledge across datasets and mitigate the inter-dataset information gap by effectively transferring the domain-agnostic knowledge. Our main contributions are summarised as follows:

1) We show that, like humans, domain-specific and domain-agnostic knowledge also exists in VideoQA. Also, the domain-agnostic knowledge is transferable and can bring considerable performance improvement.
2) For transfer learning experiments on VideoQA, we construct a new dataset, KnowIT-X, containing $21,412$ QA pairs, each of them annotated with knowledge.[1] The video clips are sampled from a TV show, *Friends*, which is a different domain from the standard model [8]'s with *The Big Bang Theory*.

2 Related Work

Video question answering Traditionally, temporal visual information and language are processed with deep neural network (DNN)-based models and fused to predict the correct answer. For example, LSTMs [12] are used to embed the temporal visual and textual features in [15], and attention mechanisms [24], which allows to focus only on specific parts of the input, bring significant improvements in [6, 30, 34, 35]. Other work [32, 33] applies Transformers [5] to capture the information from videos. For better fusion of independent visual and textual sources, Hirota *et al.* [10, 11] propose to use the textual representations to understand the visual sources. Besides the independent processing of vision and language, other approaches [2, 14] model the relationships between objects. Recent works like [31] also focus on the VideoQA without much human annotation.

Another direction is to retrieve external information from knowledge bases (KB), which successfully extends the linguistic features and steers the model to a specific part of the visual

---

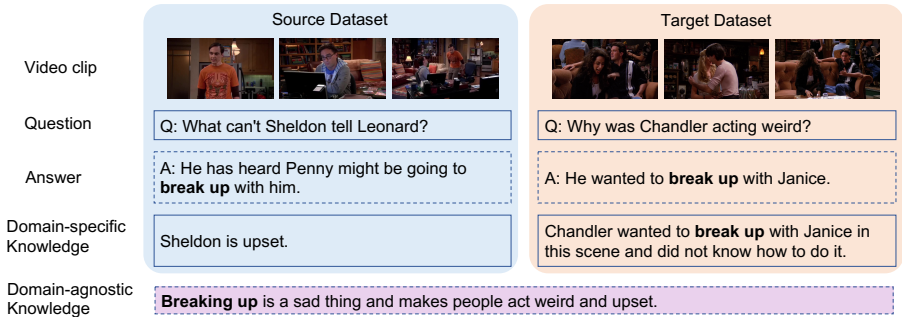[1]Available at https://knowit-vqa.github.io/

Figure 1: Examples of transfer learning on different domains of VideoQA. The two datasets (source and target) have different domain-specific knowledge but share the same domain-agnostic knowledge. We argue that the latter can be transferred across datasets.

content. For example, in visual question answering (VQA), a structured KB like ConceptNet [23] or Freebase [1] is adopted as extra inputs in [25, 26, 27] and a generic unstructured KB is investigated in [16]. For VideaQA, KnowIT VQA [8] is the first unstructured video-based dataset built by humans. ROLL [7] leverages online knowledge to answer questions about video stories, showing the great potential of knowledge-based models in VideoQA.

Transfer learning in vision and language Transfer learning has been rarely investigated in VideoQA, but it has been addressed in the related VQA tasks. Hu *et al*. [13] study transfer learning between open-ended VQA datasets by a probabilistic model. Xu *et al*. [29] learn and make use of joint features across different modalities. Chao *et al*. [3] propose a domain adaptation method to deal with the cross-dataset mismatch on images, questions, or answers. However, these approaches are not directly applicable to VideoQA due to the much richer information contained in videos. In addition, more emphasis has been put on processing the dataset itself, rather than studying the knowledge learnt by the model, which gives a more general solution to diverse domains of datasets. Therefore, this paper proposes a novel knowledge-oriented transfer learning method for VideoQA. To the extent of our knowledge, this is the first work that studies the transfer of knowledge in VideoQA.

3 Methodology

Figure 2 outlines the training pipeline of our proposed transfer learning method. First, in the *pre-training* stage, we use a source dataset to train a knowledge retrieval module of a standard knowledge-based VideoQA model (Section 3.1). In the knowledge retrieval module, a question and its candidate answers are used to retrieve the most relevant piece of knowledge from a KB. We design the pre-training stage so that the knowledge retrieval module learns to mitigate the domain-specific knowledge and focus on the domain-agnostic knowledge by pre-processing the source dataset with a domain-specific entity tagger (DET) (Section 3.2). Then, in the *transfer learning* stage, we pre-process the target dataset with DET and incorporate a data augmentation module (DA) (Section 3.3). Using the domain-agnostic knowledge acquired in the pre-training stage as a medium, we finetune the pre-trained knowledge retrieval module with the pre-processed target samples (Section 3.4). Finally, the output of
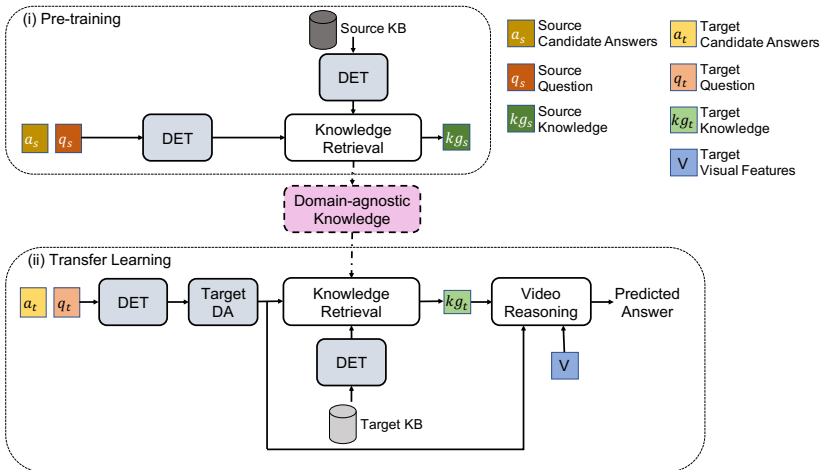
Figure 2: Overview of our transfer learning method for knowledge-based VideoQA. In (i) pre-training, the knowledge retrieval module acquires domain-agnostic knowledge. In (ii) transfer learning, this knowledge is leveraged to improve the retrieval on the target dataset.

the knowledge retrieval module, together with the visual features, question, and candidate answers, is used in a video reasoning module to predict the correct answer (Section 3.5).

3.1 Knowledge-based VideoQA model

We use ROCK [8] as the backbone VideoQA model. ROCK differs from standard VideoQA models [4, 14, 15] in that it explicitly uses external knowledge obtained from a KB, making it easier to study the problem of transferring knowledge across different domains. ROCK takes a video clip, a question, and a set of candidate answers as input, and 1) retrieves relevant knowledge to answer the question from a KB and 2) predicts the correct answer by leveraging the input and the retrieved knowledge. Specifically, ROCK presents a three-stage pipeline, including (a) KB construction, (b) knowledge retrieval, and (c) video reasoning. The KB is constructed with annotated knowledge from the original dataset [8]. The knowledge retrieval module is a BERT network [5] trained to match the most relevant knowledge in the KB given a question and candidate answers. Finally, the video reasoning module fuses the retrieved knowledge with the rest of the language and visual inputs to predict the correct answer. For further details, we refer curious readers to [8].

3.2 Domain-specific entity tagger (DET)

VideoQA models usually have strong in-domain generalisation across two sets of QA pairs with identical distributions [15]. However, in a transfer learning setup, a model is pre-trained on a source dataset and evaluated on a target dataset, where these datasets are from different origins and present different distributions. This leads to a large gap between the knowledge contained in the two datasets in terms of scenes, plots, and vocabulary. Thus, the domain-specific knowledge is specialised for a certain dataset (or domain) and may hinder the transfer to out-of-domain data [17]. To mitigate the inter-domain difference, especially in

the language modality, and facilitate more effective learning of domain-agnostic knowledge, we propose a domain-specific entity tagger (DET) to recognise and tag the domain-specific knowledge with the named entity [20] they belong to. As the entity names are more general expressions, we relieve the influence brought by the domain-specific knowledge in the source dataset so that the model does not need to know them. In this way, the transfer of domain-agnostic knowledge will be much easier by recognising the domain-specific knowledge.

Specifically, for a question q in the dataset, such as '*Why was Chandler acting weird?*', we recognise and tag the entity *Chandler* as `Person`. To maintain the original semantics and the grammatical structure, the tagged entity (*e.g.*, *Chandler*) and its entity type (*e.g.*, `Person`) are inserted back into the original sentence in an appositive form. The question DET(q) after this entity tagging becomes: '*Why was Chandler, a person, acting weird?*'.

We apply DET to question q, candidate answers $\{a_i\}_i$, and knowledge instance k in both the source and the target dataset. We denote the set after applying this by DET($\mathcal{X}$), *i.e.*,

$$\text{DET}(\mathcal{X}) = \{(\text{DET}(q), \{\text{DET}(a_i)\}_i, \text{DET}(k) \mid (q, \{a_i\}_i, k) \in \mathcal{X}\}, \tag{1}$$

where $\mathcal{X}$ denotes any VideoQA set.

3.3 Target data augmentation (DA)

When applying transfer learning from a source to a target dataset, the latter may not contain enough samples to learn sufficient domain-specific knowledge in the target domain, which may cause overfitting. To address this problem, we augment the training set $\mathcal{T}$ in the target dataset before transferring the pre-trained model. To augment the training set we propose to apply back translation [11, 22], which translates a natural language sentence into a pivot language with the function Translate($\cdot$), and then translates it back to the original language with the function BackTranslate($\cdot$). With this technique, we incorporate twice as many samples as the original dataset, with a different structure but maintaining the original semantics. Also, for extremely small-scale datasets, back translation can be applied several times using multiple pivot languages.

Formally, for any text s (*i.e.*, a question, an candidate answer, or a knowledge instance) in the training set,

$$\begin{aligned} s_{\text{PV}} &= \text{Translate}(s) \\ s' &= \text{BackTranslate}(s_{\text{PV}}), \end{aligned} \tag{2}$$

where s_{PV} is the sample in the pivot language and s' the augmented sample. As some s'_t can be almost the same as its original ($s'_t \approx s_t$), we identify them with a pre-trained BERT network [5] and similarity threshold α to remove them. We incorporate the new s'_t to the target training set $\mathcal{T}$, creating DA($\mathcal{T}$) = $\mathcal{T} \cup \mathcal{T}_{\text{BK}}$, where $\mathcal{T}_{\text{BK}}$ is the set of back-translated triplets of a question, a set of candidate answers, and a knowledge instance, after removal. We apply DA to all the questions and candidate answers in the target training set.

3.4 Domain-agnostic knowledge transfer for knowledge retrieval

With the samples tagged with domain-specific knowledge and augmented by back translation, we next introduce the transfer of the domain-agnostic knowledge. Our VideoQA model is pre-trained on the training set $\mathcal{S}$ of the source dataset and then transferred to the target dataset. During pre-training, the knowledge retrieval module is trained to locate the most

relevant knowledge instances kg in the source KB given the input source question q and source candidate answers $\{a\}$. Using $p_\theta(kg \mid q, \{a\})$, which denotes the probability of kg being relevant to q and $\{a\}$ parameterized by θ, this can be formulated as a maximisation problem of the expected log-likelihood:

$$\theta_{\text{Pre}} = \arg\max_\theta \mathbb{E}[\log(p_\theta(kg \mid q, \{a\}))], \qquad (3)$$

where the expectation is computed over all $(q, \{a\}, kg) \in \text{DET}(\mathcal{S})$. In the transfer learning stage, the pre-trained knowledge retrieval module is finetuned to find parameter θ_{FT} with the same training strategy as Eq. (3) but the expectation is computed over all $(q, \{a\}, kg) \in \text{DA}(\text{DET}(\mathcal{T}))$ and the parameters are initialised by θ_{Pre}.

3.5 Answer prediction on the target domain

To predict the final answer, we extract visual features v from the input video clip in the target dataset and fuse them in the video reasoning module together with question q, candidate answers $\{a_i\}_i$, and retrieved knowledge $\hat{kg} = \arg\max_{kg} p_{\theta_{\text{FT}}}(kg \mid q, \{a\})$ by the transferred knowledge retrieval module. We extract v at three different scales of information: image features from video frames using ResNet50 [9], facial features with recogniser [13], and captions generated with [28]. The textual data (*i.e.*, q, $\{a_i\}_i$, kg) is concatenated and encoded in single vector u_i per candidate answer for $i = 1, \ldots, N_a$ (N_a is the number of candidate answers), as in [8]. The visual and language features are concatenated and projected with a fully-connected layer to obtain an answer score. The predicted answer is the one with the highest score. The video reasoning module is trained with a multi-class cross-entropy loss.

4 KnowIT-X VQA Dataset

We use KnowIT VQA [8] as source dataset, as it contains annotated knowledge for each sample in VideoQA. As there are no other comparable datasets with knowledge for VideoQA to use as target, we collect a new one following KnowIT VQA framework. We name it KnowIT-X VQA (X for transfer learning). While KnowIT is collected from the TV show *The Big Bang Theory*, KnowIT-X is from another popular TV show: *Friends*. We choose *Friends* because it shares strong similarities with the original KnowIT VQA dataset in terms of format, video length, and audience, whereas it is different in terms of plots and topics, making it ideal for transfer learning. Additionally, to study transfer learning at multiple scales (transfer learning is commonly applied to small-scale datasets), we create three different versions: KnowIT-X full, KnowIT-X v-5k, and KnowIT-X v-3k, with all the collected samples, $5,092$ samples, and $2,900$ samples, respectively. Details are shown in Table 1. Some examples of KnowIT-X are shown in the supplementary material.

Video clip extraction   We obtain videos and subtitles from the original DVDs. We use 202 episodes, distributed in all 10 seasons of *Friends*. We divide each episode into 20-second clips by scene, where each scene is aligned with episode transcripts available online.[2] We acquire $12,176$ clips in total.

Annotations   Following KnowIT, for each video clip, we ask workers on Amazon Mechanical Turk[3] to write a question, four candidate answers (one correct), and the associated

---

[2]https://fangj.github.io/friends/
[3]https://www.mturk.com

Table 1: Comparison between original KnowIT and KnowIT-X datasets. The three versions (v) of KnowIT-X are designed to study transfer learning on knowledge-based VideoQA.

| | KnowIT [8] | KnowIT-X | | |
		v-3k	v-5k	v-Full
Num. Episodes	207	25	50	202
Num. Scenes	2,472	311	613	2,565
Num. Clips	12,078	1,535	2,989	12,176
Num. Samples	24,282	2,900	5,092	21,412
Ave. Length of Questions	7.49	7.69	7.83	7.65
Ave. Length of Substitles	56.79	38.48	39.06	39.06
Ave. Length of Wrong Answers	4.13	1.99	2.10	2.06
Ave. Length of Correct Answers	4.54	2.21	2.29	2.30
Ave. Length of Knowledge	10.39	14.37	15.00	14.44

Figure 3: Vocabulary clouds for KnowIT and KnowIT-X VQA datasets.

knowledge that makes the answer correct (see examples in Figure 1). After two sweeps of all video clips, we obtain 21,412 samples, which we split into 17,583 for training, 1,748 for validation, and 2,081 for test sets. For the smaller versions v-5k and v-3k, we randomly reduce the training set to 4,269 samples and 2,432 samples, respectively. The same reduction is also applied to the test set and the validation set. The annotation process took us 2 months.

Domain comparison While KnowIT is from the show *The Big Bang Theory* with scientific topics, KnowIT-X is focused more on the relationships among characters. This difference can significantly influence the audience's impressions and interests, causing a distribution gap between the annotations. To show this gap, we first compare the vocabularies. As illustrated in Figure 3, the vocabularies differ a lot. Additionally, we present the distributions of question types in Figure 4, where the type is defined as the first word of each question. With more 'who' questions, KnowIT-X is more focused on the vision than KnowIT, which is predominately focused on knowledge, with a larger amount of 'why' questions. More comparisons between KnowIT and KnowIT-X are presented in the supplementary material.

5 Experiments

We use PyTorch [19] to implement our models. The hyperparameters are set following [8]. In the DET module, the specific entity names include 18 different entity types.[4] The DA module employs German as the pivot language and $\alpha = 0.998$.

Evaluation metrics Following [8], we report results both on the knowledge retrieval and the video reasoning modules. Knowledge retrieval results are reported as recall at k (R@k) with $k = 1, 5, 10$, and median rank (MR). Video reasoning results are reported in terms of

---

[4]Detailed entity types can be found in https://spacy.io/models/en

Table 2: Domain-agnostic knowledge transfer results on the KnowIT-X v-Full.

	Source	Target	Learning	R@1	R@5	R@10	MR
1	KnowIT	-	Direct	0.178	0.363	0.444	17
2	KnowIT-X	-	Direct	0.303	0.528	0.618	5
3	Both	-	Direct	0.316	0.531	0.621	4
4	KnowIT	KnowIT-X	Transfer (w/o DET)	0.299	0.529	0.614	5
5	KnowIT	KnowIT-X	Transfer (w/ DET)	0.303	0.531	0.630	4

Table 3: Knowledge retrieval module results on KnowIT-X v-3k and v-5k. Each training configuration is computed with and without DET and DA.

Source	Target	Learning	v-3k		v-5k	
			R@5	MR	R@5	MR
KnowIT	-	Direct	0.010	7,864	0.010	7,798
		w/ DET+DA	0.421	13	0.397	18
KnowIT-X	-	Direct	0.472	8	0.490	6
		w/ DET+DA	0.497	7	0.490	6
Both	-	Direct	0.508	**5**	0.524	5
		w/ DET+DA	0.487	6	0.497	6
KnowIT	KnowIT-X	Transfer	0.482	6	0.526	5
		w/ DET+DA	**0.521**	**5**	**0.537**	**4**

accuracy. For the knowledge retrieval, we use MR as our primary metric and R@5 as the secondary metric, as the top-5 retrieved sentences are the ones used in the video reasoning module. Additionally, note that as MR and recall do not consider the potential similarities between other highly related samples in the knowledge base, the evaluation of the knowledge retrieval may not be representative of the real improvement on the final VideoQA task.

Domain-agnostic knowledge transfer with DET   First, we evaluate the domain-agnostic transfer learning on the knowledge retrieval module. Table 2 shows results on KnowIT-X v-Full. We compare different learning methods: direct learning on a single dataset (KnowIT, KnowIT-X, or both), and transfer learning with and without our proposed DET. Direct training on the original KnowIT (row 1) shows the worst performance by a large margin, confirming the big domain gap between the two datasets. Transfer learning without DET (row 4), which is equivalent to standard finetuning, does not improve results over direct training (row 2), verifying that in this case the model is not able to discern between domain-specific and domain-agnostic knowledge. However, when we incorporate our proposed DET (row 5), results are successfully improved, obtaining the best R@5 and MR. Direct training on both datasets (row 3) performs similar to our method, but it requires more time and memory.[5]

Small-scale knowledge transfer   The benefits of using the proposed transfer learning are even more obvious on the smaller versions of the dataset. Results on KnowIT-X v-3k and v-5k are reported in Table 3. We apply DET and DA to all the training configurations, obtaining consistent improvements and indicating that the proposed method captures the shared knowledge between different datasets. The only exception is when we train directly on both datasets (KnowIT and KnowIt-X), as the model can obtain enough information at training time at the expense of a higher computation cost. Even so, our proposed transfer learning method obtains the best performance both in KnowIT-X v-3k and KnowIT-X v-5k.

---

[5]This is because the knowledge retrieval module behaves as $O(N^2)$, where N is the number of training samples.

DET	R@1	R@5	R@10	MR
Mask-out	0.241	0.442	0.534	8
Hyphen sep.	0.294	0.530	0.633	4
Appositive	**0.303**	**0.531**	**0.630**	**4**

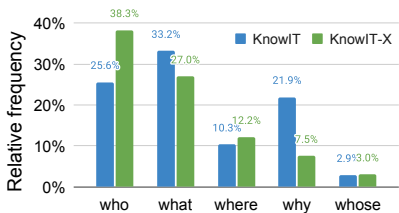Table 4: Knowledge retrieval results using different DET methods.

Figure 4: The distributions of question types in KnowIT and KnowIT-X VQA dataset.

Table 5: Video reasoning accuracy with direct learning (w/o transfer learning) on KnowIT and KnowIT-X datasets, using different combination of inputs (vision, language, and knowledge). Image features are used as the vision source in this experiment.

	Vision	Language	Knowledge	KnowIT [8]	KnowIT-X v-3k	v-5k	v-Full
1	-	QA	-	0.530	0.376	0.333	0.482
2	-	QA, Subs	-	0.587	0.497	0.515	0.622
3	✓	QA, Subs	-	0.587	0.497	0.516	0.623
4	✓	QA, Subs	Retrieved	0.652	0.579	0.663	0.738
5	-	QA, Subs	GT	0.731	0.720	0.787	0.845

Results with other metrics are presented in the supplementary material.

DET analysis We compare different ways of inserting the recognised named entities with DET and show that the appositive tagging is better than other tagging methods. In particular, we evaluate *Mask-out*, in which the detected entity is removed and replaced by its label (e.g. *Why was person acting weird?*), and *Hyphen separator*, in which the entity and its label are separated by the character "-" (e.g. *Why was Chandler-person, acting weird?*). Results are reported Table 4 for the model pre-trained on KnowIT and transferred to KnowIT-X v-Full.

Video reasoning direct results We evaluate the standard ROCK model [8] (*i.e.*, without transfer learning) on the three versions of KnowIT-X and compare the results with the original KnowIT dataset. We use image features for the vision; QA, and subtitles (Subs) for the language; and retrieved and ground truth knowledge for the KB. All the models are trained from scratch. Accuracy is shown in Table 5. We observe that 1) the accuracy when only using questions and answers (row 1) is lower on KnowIT-X, indicating that the language bias is not as strong in KnowIT-X, 2) the small-scale datasets perform significantly worse than its full counterpart, indicating the importance of training data. 3) the performances with retrieved knowledge (row 4) surpass the one without knowledge (row 3) by a large margin, which demonstrates that the retrieved knowledge predicts the correct answers.

Video reasoning transfer results In Table 6, we show the video reasoning results on the KnowIT-X v-3k, v-5k and v-full with different scales of visual information. We compare the full transfer learning method against variations with and without DET and DA, and the direct learning method. It can be seen that each strategy contributes to improving the final VideoQA accuracy. When both DET and DA are included, the result is the best. Also, among all the vision sources, captions usually help acquire the best performance. Additionally, in Figure 5, we show a qualitative comparison between direct and transfer learning. In direct learning, the retrieved knowledge is too much focused on the domain-specific entity 'hat' and has limited

Table 6: Video reasoning accuracy with the proposed transfer learning methods on KnowIT-X v-3k, v-5k and v-full. Transfer is pre-trained on KnowIT.

Vision	Learning	Knowledge	DET	DA	KnowIT-X v-3k	KnowIT-X v-5k	KnowIT-X v-full
Image	Direct	Retrieved	-	-	0.579	0.663	0.738
Image	Transfer	Retrieved	-	-	0.584	0.678	0.721
Image	Transfer	Retrieved	-	✓	0.629	0.698	0.731
Image	Transfer	Retrieved	✓	-	0.660	0.692	0.726
Image	Transfer	Retrieved	✓	✓	0.665	0.717	0.740
Caption	Direct	Retrieved	-	-	0.624	0.669	0.756
Caption	Transfer	Retrieved	-	-	0.629	0.675	0.734
Caption	Transfer	Retrieved	-	✓	0.645	0.701	0.747
Caption	Transfer	Retrieved	✓	-	0.670	0.712	0.739
Caption	Transfer	Retrieved	✓	✓	0.690	0.723	0.758
Facial	Direct	Retrieved	-	-	0.579	0.663	0.739
Facial	Transfer	Retrieved	-	-	0.584	0.678	0.721
Facial	Transfer	Retrieved	-	✓	0.629	0.698	0.731
Facial	Transfer	Retrieved	✓	-	0.660	0.692	0.726
Facial	Transfer	Retrieved	✓	✓	0.665	0.717	0.740
-	-	GT	✓	✓	0.802	0.821	0.850

How does Phoebe know the man wearing the hat?
A1: They used to attend art class together.
A2: They used to mug people on the streets.
A3: They used to play in the same band.
A4: They used to go to the same high school.

Direct Training
Retrieved Knowledge: From Joey meeting the Duchess of York and Richard Branson (yep, that's him selling Union Jack hats), to Phoebe yelling at Jennifer Saunders (yup, that was Emily's mum), it was 100% British.
A4: They used to go to the same high school.

Transfer Learning (Ours)
Retrieved Knowledge: Phoebe, a person, tells Ross, a person, that she used to know him from the streets and they mugged people
A2: They used to mug people on the streets.

Figure 5: Example results: (left) direct learning, (right) proposed transfer learning.

contributions to the final answer. On the contrary, transfer learning successfully recognises it as a domain-specific entity, putting more emphasis on the relationship between people. The retrieved knowledge is highly connected with the question and the correct answer.

6 Conclusion

We proposed a knowledge-based transfer learning framework for VideoQA. We divided the knowledge learned by the VideoQA models into two categories and achieved better transfer results by mitigating the influence of the domain-specific knowledge and transferring the domain-agnostic knowledge. Back translation was applied to augment small-scale datasets. By constructing a new knowledge-based VideoQA dataset, the experimental results showed that domain-agnostic knowledge can be transferred adaptively from the source dataset and boost performance in the proposed framework.

References

[1] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Free-base: A collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250, 2008.

[2] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. In *WACV*, pages 1–13, 2021.

[3] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *CVPR*, pages 5716–5725, 2018.

[4] Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Hierarchical relational attention for video question answering. In *ICIP*, pages 599–603, 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[6] Deniz Engin, François Schnitzler, Ngoc QK Duong, and Yannis Avrithis. On the hidden treasure of dialog in video question answering. In *ICCV*, pages 2064–2073, 2021.

[7] Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. In *ECCV*, pages 581–598, 2020.

[8] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. KnowIT VQA: Answering knowledge-based questions about videos. In *AAAI*, pages 10826–10834, 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Yusuke Hirota, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, Ittetsu Taniguchi, and Takao Onoye. A picture may be worth a hundred words for visual question answering. *arXiv preprint arXiv:2106.13445*, 2021.

[11] Yusuke Hirota, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, Ittetsu Taniguchi, and Takao Onoye. Visual question answering with textual representations for images. In *ICCV workshop*, pages 3154–3157, 2021.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[13] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In *CVPR*, pages 5428–5436, 2018.

[14] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, pages 11021–11028, 2020.

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018.

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019.

[17] Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, and Bohyung Han. Transfer learning via unsupervised task discovery for visual question answering. In *CVPR*, pages 8385–8394, 2019.

[18] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, pages 41.1–41.12, 2015.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035. 2019.

[20] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A Python natural language processing toolkit for many human languages. In *ACL*, pages 101–108, 2020.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2016.

[22] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96, 2016.

[23] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptnNet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451, 2017.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[25] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. FVQA: Fact-based visual question answering. *IEEE TPAMI*, 40(10):2413–2427, 2017.

[26] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, pages 1290–1296, 2017.

[27] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, pages 4622–4630, 2016.

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[29] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multi-modal domain adaptation. In *EMNLP*, pages 367–376, 2020.

[30] Hongyang Xue, Zhou Zhao, and Deng Cai. Unifying the video and question attentions for open-ended video question answering. *IEEE TIP*, 26(12):5656–5666, 2017.

[31] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021.

[32] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. BERT representations for video question answering. In *WACV*, pages 1556–1565, 2020.

[33] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. A comparative study of language transformers for video question answering. *Neurocomputing*, 445:121–133, 2021.

[34] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. In *SIGIR*, pages 829–832, 2017.

[35] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, pages 3518–3524, 2017.