Worldly Wise (WoW) - Cross-Lingual Knowledge Fusion for Fact-based Visual Spoken-Question Answering

Kiran Ramnath[1], Leda Sari[1], Mark Hasegawa-Johnson[1], and Chang Yoo[2]

[1]University of Illinois, Urbana-Champaign
[2]KAIST
[1]{kiranr2, lsari2, jhasegaw}@illinois.edu
[2]{cd_yoo}@kaist.ac.kr

Abstract

Although Question-Answering has long been of research interest, its accessibility to users through a speech interface and its support to multiple languages have not been addressed in prior studies. Towards these ends, we present a new task and a synthetically-generated dataset to do Fact-based Visual Spoken-Question Answering (FVSQA). FVSQA is based on the FVQA dataset, which requires a system to retrieve an entity from Knowledge Graphs (KGs) to answer a question about an image. In FVSQA, the question is spoken rather than typed. Three sub-tasks are proposed: (1) speech-to-text based, (2) end-to-end, without speech-to-text as an intermediate component, and (3) cross-lingual, in which the question is spoken in a language different from that in which the KG is recorded. The end-to-end and cross-lingual tasks are the first to require world knowledge from a multi-relational KG as a differentiable layer in an end-to-end spoken language understanding task, hence the proposed reference implementation is called Worldly-Wise (WoW). WoW is shown to perform end-to-end cross-lingual FVSQA at same levels of accuracy across 3 languages - English, Hindi, and Turkish.

1 Introduction

Imagine being able to ask your voice assistant a question in any language, to learn some trivia about your favorite movie star. This task falls in the realm of Knowledge-based Question Answering (QA). One such challenging QA task is that of Fact-based Visual Question Answering (FVQA) (Wang et al., 2018) which seeks to imitate how humans leverage background common-sense knowledge when answering visual questions. This task ensures that answering each question about an image requires external knowledge not directly available within the image or the text of the question. (see Fig. 1). The external information is provided in the form of knowledge graphs, which are multi-relational graphs, storing relational representations between entities. The entities could be single words or phrases of words that denote objects or concepts. Such tasks, though widely studied, exist mostly for well-resourced languages (Goyal et al., 2017; Wang et al., 2018). These languages generally also have mature Automatic Speech Recognition (ASR) systems and language models. The accompanying Knowledge Graphs (KGs) also tend to be limited to languages that are well-resourced (Auer et al., 2007; Tandon et al., 2014; Liu and Singh, 2004). Against this background, it is worthwhile to think of building end-to-end systems which directly use speech signals as input, that can readily harness huge knowledge repositories stored in another language, instead of requiring Tabula Rasa learning.

Figure 1: Example of a fact-based visual question
Question - Which object in this image can be found in a Jazz Club?
Supporting fact - You are likely to find [[a trumpet]] in [[a jazz club]]
Subject, Predicate, Object - (Trumpet, AtLocation, Jazz Club)
Answer - Trumpet

With these motivations, the main contributions of this paper are two-fold: 1) A new task referred to as Fact-based Visual Spoken-Question Answer-

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1908–1919
June 6–11, 2021.

ing (FVSQA) along with the release of 5 hours of synthetic-speech data in each of the three languages - English, Hindi, and Turkish. 2) An end-to-end architecture Worldly-Wise (WoW) capable of answering questions trained directly on speech features in all three languages. To the best of our knowledge, this is the first work to perform KG knowledge acquisition using only a speech signal as input, without the requirement for a pre-trained automatic speech recognizer as a system component.

Worldly-Wise (WoW) is readily generalizable to other languages, even those without an ASR-system. This is possible because of two reasons - a) it obtains speech features as Mel-Frequency Cepstral Coefficients and does not require ASR-based text-conversion or speech feature extraction from a language-specific pretrained network, and b) for knowledge acquisition, it does not require the entity label to be in the same language as the question, instead leveraging neuro-symbolic entity representations in the form of KG embeddings. These KG embedding methods, trained to remedy KG sparsity by performing missing-edge prediction, learn transferable entity-features that encode the local and global structures in KGs. This also permits the architecture to use an image representation technique called 'Image-as-Knowledge' (IaK). This uses a co-attention mechanism that attends to important entities in the image and time-steps in a question, thus allowing for improved answer retrieval. The IaK technique was first presented by (Ramnath and Hasegawa-Johnson, 2020) for the goal of performing FVQA over incomplete KGs, but is applied to a speech signal as opposed to a textual question. We revisit its important details below in the relevant sections.

We report experimental results on synthetic speech data in the aforementioned diverse languages to demonstrate its effectiveness. Hindi and Turkish are simulated as under-resourced languages by denying the system access to any text, ASR, or machine translation to or from those languages, thereby requiring the system to learn the mapping from Hindi and Turkish speech signals to the KG knowledge stored in English. Through this work, we hope to motivate research in expanding spoken language understanding (SLU) in under-resourced languages through models which circumvent the need for parallel text labelled resources.

2 Related Work: Multimodal SLU

Spoken language understanding (SLU) has a long history. It is well established in speech literature that using speech audio features in an end-to-end fashion for Language Understanding tasks is non-trivial compared to text. There are several difficulties in using speech directly as input such as long length of inputs making it difficult to densely capture context, presence of spoken accents, gender, environmental noise, and acoustic information, etc. which all pose challenges for use in end-to-end semantic reasoning on it.

For most of its history, SLU was developed in a pipelined fashion, with ASR feeding text to a natural language understanding system, e.g., to the best of our knowledge, the only published uses of SLU with knowledge graphs that fit this description is (Woods, 1975). Recent research in end-to-end multimodal SLU bypasses the need for ASR by leveraging a parallel modality such as image (Harwath et al., 2016; Kamper et al., 2019) or video (Sanabria et al., 2018), or a non-parallel corpus of text (Sarı et al., 2020), to guide learning speech embeddings such that the speech input can be used in a downstream task.

In speech-based VQA applications, the most common approach is a two-step approach which consists of an ASR followed by text-based VQA (Zhang et al., 2017). However, these systems are not generalizable to under-resourced or unwritten languages for which we cannot train an ASR system. Therefore, in this study, we will explore using neural speech embeddings, which are guided by the information in the KG, for achieving FVSQA.

3 Related Work: Knowledge Graphs

Knowledge graphs (Suchanek et al., 2007; Auer et al., 2007; Bollacker et al., 2008) are effective ways of representing objects or concepts and their inter-relationships. Such relational representations are formally defined in the Resource Description Framework (RDF) as triples $f = (subject, predicate, object)$, where $(subject, object)$ are entities, $predicate$ is the relation connecting the two entities. (Halford et al., 2010) showed that such linked representations correlate highly with human cognition. Furthermore, KGs can be classified as Closed-World or Open-World. The former assumes that non-existent fact triples must necessarily be false, while the latter assumes that the KG could be incomplete, and there-

fore missing edges could be either true or false. While closed-world assumptions hold for domain-specific KGs, common-sense KGs extracted from web-scale datasets do not respect this assumption (Galárraga et al., 2013; Dong et al., 2014).

3.1 KG embeddings

Common-sense KGs extracted from web-scale datasets are usually incomplete. KG embedding techniques (Bordes et al., 2013; Sun et al., 2019; Socher et al., 2013; Nickel et al., 2011; Dong et al., 2014; Dettmers et al., 2018) have been studied as a means to remedy incompleteness of large-scale KGs. These embeddings have been shown to transfer well to other tasks that require knowledge acquisition over the KGs.

KG Embedding methods usually assign scores or truth-probabilities to each fact triple by learning latent features for entities and relationships. These methods learn a score mapping $\phi(h, r, t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbf{R}$ where $\mathcal{E}$ is the set of all entities, $\mathcal{R}$ is the set of all relation-types. $h, t \in \mathcal{E}$ are the head (subject) and tail (object), $r \in \mathcal{R}$ is the directed relationship that connects the two. The observed KG can be expressed as $\mathcal{G} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, which in turn is a subset of $\mathcal{G}_o$, the unknown set of all true edges in the world that the KG seeks to represent. The embeddings (h, r, t) are learned so that the score $\phi(.)$ is high for edges not just in $\mathcal{G}$ but also for those in $\mathcal{G}_o$, and low for edges outside of it.

Distance-based models (Bordes et al., 2013; Sun et al., 2019; Trouillon et al., 2016; Bordes et al., 2011) learn embeddings h, r and t in order to minimize the distance between t and $f(h, r)$, for some projection function $f(\cdot)$. Common-sense KGs are often based on free text, therefore most entities occur rarely; an example is the entity "lying on" in Fig. 2. Since it is very challenging for distance-based methods to perform completion of common-sense KGs, very few previous benchmarks have approached this task (Li et al., 2016; Malaviya et al., 2020). In (Ramnath and Hasegawa-Johnson, 2020), it was shown that Entity-Relation Multi-Layer Perceptron (ERMLP) (Dong et al., 2014), which uses an MLP to produce the score $\phi(h, r, t)$ for each fact triple, works better for FVQA in comparison to TransE and RotatE.

3.2 KGQA

Knowledge-graph question answering (KGQA) is the task of answering questions regarding facts that can be inferred/retrieved from a KG given the question, image and the graph. Language-only benchmarks include (Bordes et al., 2015; Berant et al., 2013), vision-and-language benchmarks include (Sanket Shah and Talukdar, 2019; Marino et al., 2019; Wang et al., 2018). In (Wang et al., 2018), FVQA is approached as a parsing and fact-retrieval problem, while (Narasimhan and Schwing, 2018) directly retrieves facts using lexical-semantic word embeddings. In Out-of-the-box (OOB) reasoning (Narasimhan et al., 2018), a Graph Convolutional Network (Kipf and Welling, 2017) is used to reason about the correct entity, while (Zhu et al., 2020) (the current State-of-the-Art in the complete-KG FVQA task) added a visual scene-graph (Krishna et al., 2016) and a semantic graph based on the question alongside the (OOB) KG reasoning module. In (Ramnath and Hasegawa-Johnson, 2020), FVQA is tackled on incomplete KGs using KG embeddings to represent entities instead of word-embeddings, as the latter are shown to be inadequate for this task.

Among other KGQA works closely related to our approach, (Huang et al., 2019) answer a text question using minimum-distance retrieval of translational KG entity and relation embeddings, thereby achieving SOTA results on SimpleQuestions with supporting knowledge bases Freebase2M and Freebase5M (Bollacker et al., 2008). In (Lukovnikov et al., 2017), authors use character-level embeddings for SimpleQuestions. In (Saxena et al., 2020), KG Embedding-based reasoning over missing edges is performed on the text-only benchmarks Webquestions (Berant et al., 2013) and MetaQA (Zhang et al., 2018), where they also perform multi-hop reasoning. Amongst KGQA baselines involving the visual modality, the OKVQA benchmark (Marino et al., 2019) provides outside common-sense knowledge in the form of supporting text. The accompanying external knowledge is acquired using a neural network parse of the fact text. KVQA (Sanket Shah and Talukdar, 2019) provided KGs as outside knowledge, and they tackled the task using face-recognition and entity-linking to answer several different types of questions.

4 Task Formulation

This section introduces a new task called FVSQA and presents a new dataset collected for this task.

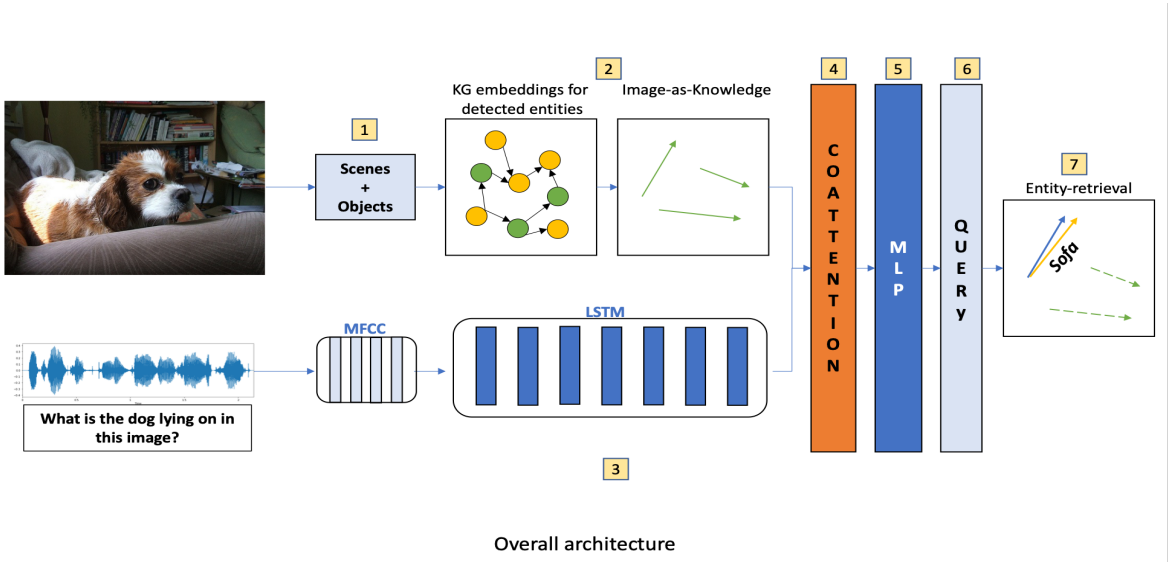


Figure 2: Our architecture for FVSQA. (1) Object and scene detectors find constituent entities in images. (2) The image is represented as a collection of KG embedding features for these detected entities. (3) MFCC features for the spoken question are passed via an LSTM. (4) The co-attention mechanism described in Fig. 3 fuses the image and question encoding, then (5) passed through successive fully-connected layers, whose (6) last layer is used as a query. (7) The closest entity to this query is retrieved as the answer.

4.1 FV(S)QA

FVSQA is similar to FVQA in all aspects but for the modality of the question q; in FVSQA it is a speech input instead of a text input.

The following condition holds for questions in the FVQA (Wang et al., 2018) benchmark: for each (question,image,answer) triplet in the dataset ($(q_i, I_i, y_i) \in \mathcal{D}$), exactly one supporting fact in the knowledge graph ($f_j = (h, r, t) \in \mathcal{G}$) exists such that the correct answer y_i is either the head or the tail of f_j, and such that at least one of the two entities is visible in the image.

The companion knowledge-graph is constructed from three diverse sources: ConceptNet (Liu and Singh, 2004), Webchild (Tandon et al., 2014), and DBPedia (Auer et al., 2007). ConceptNet provides common-sense knowledge about entities, DBPedia mainly conveys hypernym (i.e. parent-child) relationships, while Webchild covers many different kinds of comparative relationships between entities (these are considered as a single relationship-type for FVQA).

Answering questions in FVQA is to perform the following operation

$$\hat{y} = \underset{e \in \mathcal{E}}{\operatorname{argmax}}\ p(y = e \mid q, I, \mathcal{G}), \qquad (1)$$

i.e., retrieving that entity which is most likely to be the correct answer given a question q and image I, and given the graph $\mathcal{G}$.

Knowledge Base	Total facts	Questions
DBPedia	35152	817
ConceptNet	119721	4652
Webchild	38576	357

Table 1: Distribution of facts and questions across the KBs (Wang et al., 2018)

The FVSQA task formulation is identical, except that the question is not textual but spoken. We study the task when the question is spoken in one of three languages – English, Hindi, Turkish.

4.2 Data Description

The dataset contains 2190 images sampled from the ILSVRC (Russakovsky et al., 2015) and the MSCOCO (Lin et al., 2014) datasets. 5826 questions were obtained via crowdsourcing on Amazon Mechanical Turk which concern 4216 unique supporting facts (Table 1). FVSQA provides the same five train-test splits as FVQA, where each split contains images and questions roughly in the ratio 1:1. The accompanying KG consists of roughly 194500 facts, about 88606 entities. In total, the dataset contains 13 relations: $R \in$ *{Category, HasProperty, RelatedTo, AtLocation, IsA, HasA, CapableOf,*

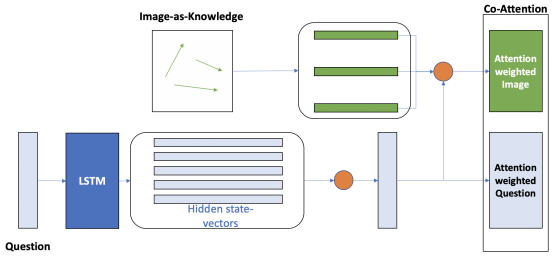


Figure 3: A co-attention mechanism fuses the image and question representations. First, self-attention provides a single question-embedding of the speech signal (bottom orange circle). Next, the question-embedding functions as a context vector to guide the visual attention weights (top orange circle). The final image embedding is a vector present in the span described by its constituent entities' KG Embedding vectors.

UsedFor, Desires, PartOf, ReceivesAction, CreatedBy, Comparative}.

The next section describes how the multi-lingual speech data is generated.

4.2.1 Data Generation - Text Translation

The text questions in FVSQA dataset are in English. To generate spoken questions in Hindi and Turkish, we first translate the questions using Amazon Translate API[1] from English. We manually review the questions to ensure intelligibility of questions. These translated texts are only used for speech data generation; these are not available to the network during either training or inference.

4.2.2 Data Generation - Text-to-Speech

We use Amazon's Polly API[2] to generate spoken questions for each language. The generated speech is in mp3 format, sampled at 22 kHz. For a given language, all questions were generated using the same voice. The voices used were Joanna for English, Aditi for Hindi, and Filiz for Turkish. We again manually review and ensure intelligibility of speech data so generated.

5 Our Approach

Fig. 2 depicts the architecture we use for FVSQA. As shown in the figure, co-attention fuses an image I and question q to form a query vector ν. This query vector is then used to retrieve the answer from the KG as

$$\hat{y}(q|I) = \underset{e \in \mathcal{E}}{\operatorname{argmax}}\, \nu(q, I)^T e. \qquad (2)$$

[1]https://aws.amazon.com/translate/
[2]https://aws.amazon.com/polly/

The following sections address representations of the question, KG, and image, the information fusion function $\nu(q, I)$, and the loss function. The image and KG representations are identical to those considered in (Ramnath and Hasegawa-Johnson, 2020), however, their goal is different from ours, as they perform monolingual text-FVQA over incomplete KGs.

5.1 Question representation

We represent the speech waveforms using Mel-Frequency Cepstral Coefficient features. We set the window-length to 25 ms and stride-size of 10 ms. For each time-step, we follow standard convention of using 39-dimensional vectors - the first 12 cepstral coefficients and the energy term, along with delta and double-delta features to gather contextual information as well.

5.2 KG Representation

To discriminate between a true and false fact, a binary classification-based KG Embedding model is used. Training a meaningful classifier would require presenting it with both positive and negative examples, but the observed KG $\mathcal{G}$ has only positive samples. This leads us to a 'chicken and egg' problem – KG Embeddings are supposed to mitigate the very problem of incompleteness, yet they need some negative edges to actually learn a good score function. Some heuristics have been empirically found to work well in overcoming this problem. Under the Locally Closed World Assumption (LCWA) (Dong et al., 2014), negative samples can be generated by randomly corrupting the tail entity of existing facts. The KG embedding loss function penalizes the network when a true edge has a low truth-probability, and a false edge has a high truth-probability. But some false facts may be more difficult for the model to classify as false than the others. (Sun et al., 2019) introduced a self-adversarial negative sampling strategy so that the loss function reflects this, and each false fact's contribution to the loss is scaled by the truth-probability assigned by the network during training. Thus, false edges with a higher truth-probability are penalized more heavily than false edges with lower truth-probabilities.

Based on each true fact f_i, a total of n adversarial facts are generated and used to train discriminative embeddings using noise contrastive estimation (Gutmann and Hyvärinen, 2010). Thus the knowledge graph embedding loss $\mathcal{L}_{KGE}$ in-

Module	No. of parameters
ERMLP	27,306,901
LSTM	12,480
Image representation	0
Visual Attention	340
Textual Attention	40
FVSQA MLP	193860

Table 2: Number of parameters in WoW

cludes the arithmetic inverse of the sum of the log probability that each observed edge is true ($\ln \sigma(\phi(f_i))$), plus the expected log probability that the adversarial edges are false $\Big(\ln \sigma(-\phi(f'_j)) = \ln\Big(1 - \sigma(\phi(f'_j))\Big)\Big)$:

$$\mathcal{L}_{KGE} = -\sum_{i=1}^{|\mathcal{G}|}\Bigg(\ln \sigma\big(\phi(f_i)\big) + \sum_{j=1}^{n} p_i(f'_j) \ln \sigma\big(-\phi(f'_j)\big)\Bigg) \quad (3)$$

where expectation is with respect to the probability $p_i(f'_j)$. This probability is tuned using a temperature hyperparameter α as

$$p_i(f'_j) = \frac{\exp(\alpha\, \phi(f'_j))}{\sum_{k=1}^{n} \exp(\alpha\, \phi(f'_k))}. \quad (4)$$

Eq. (3) is used to train embeddings of the head (h) and tail (t), which are applied to the FVSQA task as described in the next several subsections. Eq. (3) also trains relation embeddings (r) and MLP weights for the ERMLP scoring function (w_{MLP}); these quantities are not used for the downstream FVSQA task.

5.3 Image as Knowledge (IaK) Representation

We revisit the IaK representation first described by (Ramnath and Hasegawa-Johnson, 2020). For the FVQA task, (Narasimhan and Schwing, 2018) established the importance of representing images as a bag-of-visual concepts instead of using features from pretrained networks. This is a simple one-hot encoding of all object and scene detections found in the image. IaK instead represents each image as a contextually-weighted sum of KG entity vectors of detected visual concepts. (Ramnath and Hasegawa-Johnson, 2020) showed its superior performance for text-FVQA.

Detecting Objects: We use Torchvision's COCO object-detector to detect the 80 COCO (Lin et al., 2014) object classes. The detector used was a Faster RCNN network (Ren et al., 2015) with a ResNet50 backbone (He et al., 2016), and feature pyramid network (Lin et al., 2017). Another detector (ZFTurbo, 2018) trained on OpenImages 600 classes detections was used; we then retain only those classes which are present in ImageNet 200 object detection classes as well as in (Wu et al., 2016). The overlap obtained is almost exact; fewer than 10 classes were not found.

Detecting Scenes: A WideResNet (Zagoruyko and Komodakis, 2016) detector trained on the MIT365 places dataset (Zhou et al., 2017) detects the scenes depicted in each image. Only those classes which were used for constructing the FVQA KG (i.e. the 205 classes from MIT205 places dataset) are retained.

Upon detecting objects and scenes in each image, their corresponding entity KG embeddings are retrieved from KG. IaK then represents each image as a concatenation of entity embedding vectors.

More specifically, $I_i = [e_i^1, \ldots, e_i^m] \in \mathbf{R}^{N_e \times m}$, where N_e is the embedding dimension, and m is the number of visual concepts detected in the image.

5.4 Fusion Function ν

As shown in Fig. 3, a co-attention mechanism fuses the image and question representations. To compute a contextual-query for the image-attention, we first obtain a self-attention weighted question representation $A(q_i)$ as:

$$A(q_i) = \sum_{t=1}^{|q_i|} \alpha_q^t\, q_i^t, \quad \alpha_q^t = \frac{\exp(w_{\alpha_q}^T q_i^t)}{\sum_{t=1}^{|q_i|} \exp(w_{\alpha_q}^T q_i^t)}, \quad (5)$$

where α_q^t, w_{α_q} are respectively the attention paid to time-step w_t, and the weight parameters of the attention network used to compute the attention-scores.

Then, using $A(q_i)$ as a query, a contextual attention-weighted summary of the image $A(I_i)$

Method	MR	MRR	Hits@1	Hits@3	Hits@10
ERMLP	11194	0.156	0.132	0.152	0.197

Table 3: KG Embedding accuracy

Language	Hits @1	Hits @3
English	49 ± 0.62	61.85 ± 1.13
Turkish	48.96 ± 1.14	61.56 ± 0.79
Hindi	49.29 ± 0.73	61.26 ± 0.93
English - ASR + text-FVQA	54.07 ± 1.15	65.52 ± 0.75

Table 4: FVSQA Performance of WoW architecture across different languages

is obtained as:

$$A(I_i) = \sum_{j=1}^{m} \alpha_I^j \, e_i^j,$$
$$\alpha_I^j = \frac{\exp\left(w_{\alpha_I}^T \begin{bmatrix} A(q_i) \\ e_i^j \end{bmatrix}\right)}{\sum_{k=1}^{m} \exp\left(w_{\alpha_I}^T \begin{bmatrix} A(q_i), \\ e_i^k \end{bmatrix}\right)} \quad (6)$$

where $\alpha_I^j, w_{\alpha_I}, e_i^j$ are respectively the attention paid to concept j in the image, the weight parameters of the image-attention network, and the j^{th} constituent concept of the image.

$A(I_i)$ represents a mapping: $R^{N_e \times m} \rightarrow R^{N_e}$, which is the attention-weighted convex combination of its inputs, thus $A(I_i)$ is a vector drawn from the span of the entities present in the image. $A(q_i)$ represents a mapping: $R^{39 \times T} \rightarrow R^{39}$, T being the length of the spoken question signal.

Finally, a query vector is obtained by fusing the attention-weighted image and question vectors in the following manner:

$$\nu(q_i, I_i) = h\left(A(I_i), A(q_i); w_\nu\right) \quad (7)$$

where $h(\cdot)$ is a two-layer fully-connected network with ReLU activation functions. As prescribed in STTF (Narasimhan and Schwing, 2018), late fusion is used wherein both the question and image vectors are separately passed through one fully-connected layer before being concatenated.

5.5 Loss function

The loss function in Eq. 8 mirrors the answer prediction mechanism, in that the network is penalized whenever the cosine-similarity between the produced query and ground-truth answer deviates from 1.

$$\mathcal{L}_{FVQA} = \sum_i \left(1 - y_i^T \hat{y}(q_i|I_i)\right) \quad (8)$$

where $\hat{y}(q_i|I_i)$ is as given in Eq. (2).

6 Experimental Setup

Apart from the MFCC feature generation, the rest of the experimental setup is similar to that described in Seeing-is-Knowing (Ramnath and Hasegawa-Johnson, 2020). It is briefly recapped in the sections below.

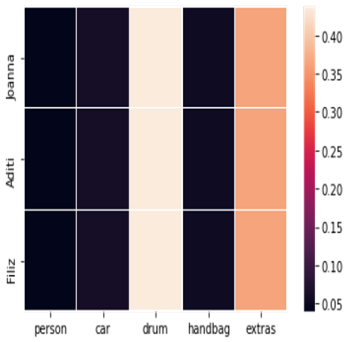
6.1 Training the KG Embeddings

For training KG Embeddings, the entire KG is split as 80% training set and 20% test set. The embedding dimensions for both entity and relation embeddings are $N_e = N_r = 300$. The batch size used is 1000. ERMLP is trained for 25,000 epochs. Adam optimizer is used for which the learning rate was initialized as 0.01 and then it is scaled down by a factor of 0.1 after every 10,000 epochs. The hyper-parameter search for the learning rate was performed by choosing among values in the set $\{0.0001, 0.001, 0.01, 0.1\}$. The temperature hyper-parameter α for the self-adversarial probability parameterization is set to 1 for all experiments. The number of adversarial samples n generated for each positive sample is 16.

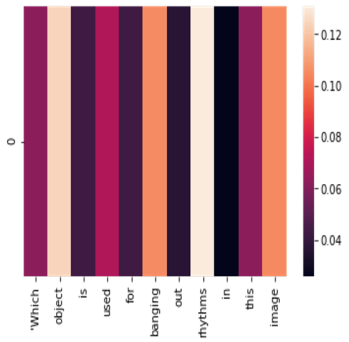
ERMLP is parameterized as a three-layer neural network. The size of the first layer is $3N_e$ since it takes the concatenated head, relation, and tail embeddings as input. Subsequent layers are $2N_e$ and N_e in size respectively, which are finally capped by a single sigmoid unit to output the truth probability $\phi(h, r, t)$. The activation functions used by the hidden layers are the Rectified Linear Unit (ReLU), which outputs $max\{0, x\}$ for an input x. All layers are fully connected and none of them use dropout.

The KG Embeddings accuracy is measured using the standard metrics: Hits @1, Hits @3, Hits @10. These determine how often each correct tail/head gets ranked in the top 1, 3, or 10 ranked facts for

Question 1: Which object is used for banging out rhythms in this image?
SPO triple: {Drum, UsedFor, banging out rhythms}
Answer Source: Image
Answer: Drum
Answer predicted: Drum

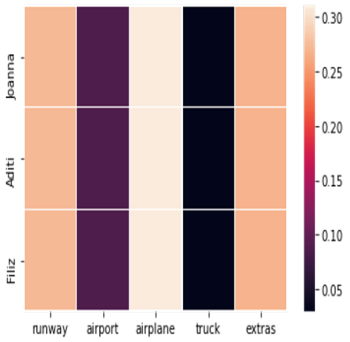


Speech attention summed over all time-time-steps of a word

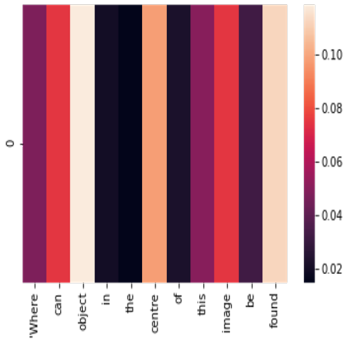


Visual Attention for all three voices

Question 2: Where can object in the center of image be found?
SPO triple: {Airplane, AtLocation, Airport}
Answer Source: Image
Answer: Airport
Answer predicted: Runway



Speech attention summed over all time-steps of a word



Visual Attention for all three voices

Figure 4: Visualizing co-attention maps produced by Worldly Wise

each ground-truth $(h, r)/(r, t)$ pair. Mean Rank is a metric often used to gauge the performance of KG Embeddings. It measures the mean rank of each true fact $f_i := (h, r, t)$ in the dataset when ranked by its truth-probability for a given (h, r) pair. An allied metric is the Mean Reciprocal Rank = $\frac{1}{|\mathcal{D}|}\Sigma_i \frac{1}{R_i}$.

6.2 Training WoW

A maximum of $m = 14$ visual concepts are detected in each image. We report Hits @1 and Hits @3 for each model. All the results are based on performing K-fold cross validation across the five train-test splits; the numbers reported are mean and standard deviation. To train the fusion function ν, the optimizer used is Stochastic Gradient Descent with a batch size of 64. The training runs for 100 epochs with a learning rate of 0.01 and a weight decay of 1e-3. Fully-connected layers use a dropout probability of 0.3.

All models were trained using GPU servers provided by Google Colab. The training for the ERMLP takes approximately 3 hours, while training $\nu(q, I)$ on one train split takes roughly 2 hours.

7 Results and Discussion

7.1 Cross-lingual FVSQA

Aided by ERMLP, WoW is able to perform FVSQA at the same levels of accuracy across English, Hindi, and Turkish. FVSQA is trained using the best performing KG embedding model demonstrated in (Ramnath and Hasegawa-Johnson, 2020) and its performance is highlighted in Table 3. To verify the superiority of ERMLP over word-embeddings, we compare a model trained with KG entities represented as averaged word embeddings instead. This representation fails to train an end-to-end system even for English, the final accuracy being close to 0%.

For English, we additionally investigate an ASR + Text-based system, where the FVQA model is trained on gold-standard textual questions, and dur-

ing inference-time, an ASR-converted speech transcript of the question is provided. The ASR system is based on the pre-trained Kaldi ASpIRE model[3] which was originally trained on augmented Fisher English dataset. The resulting FVQA system performs better than an end-to-end system for English. This indicates some joint-training strategies for speech and text-based systems could help increase accuracy for the end-to-end speech system. However, our experiments on sharing the lower layers of the network between speech and text-systems did not improve accuracy of the end-to-end speech system for English.

7.2 Attention mechanism visualizations

We can see in Q.1, Fig. 4 that for each language, the speech signal can perform as a good query vector to calculate contextual visual attention as per Eq.(5). The resulting IaK attention maps are interpretable, and in cases where the network predicts the wrong answer, provide an insight into the reason for the network's failure as in Q.2.

Furthermore, the speech self-attention maps are also coherent and informative. The alignment of time-steps in the speech signal with boundaries is generated alongside the question generation. This information, however, is not used while training the network, and is only used to investigate the attention mechanism. Fig. 4 also shows attention accumulated by each word over all time-steps of the word's utterance. We can clearly see that the relevant time-steps are attended to, depending on the image and the question itself. To the best of our knowledge, this is the first work to jointly learn attention-based speech representations guided by external KG knowledge.

8 Conclusion

A new task FVSQA is presented in this work, along with an architecture that can perform cross-lingual knowledge acquisition for question-answering. In the process, we demonstrate the first task to perform knowledge acquisition directly using a speech signal as an input. This knowledge acquisition for speech can be extended to other tasks such as audio caption-based scene identification (Harwath et al., 2016) and multi-modal word discovery (Harwath et al., 2018). Future work will include extending FVSQA to a multi-speaker setting, gathering spoken data from real-world speakers, as well as extending it to languages without an ASR system.

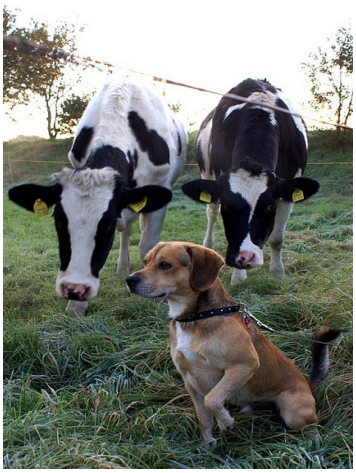


Figure 5: Example of a fact-based visual question
Question - Which animal in this image is man's best friend?
Supporting fact - [[dogs]] are [[man's best friend]]
Subject, Predicate, Object - (Dog, HasProperty, man's best friend)
Answer - Dog

9 Ethical Impact

We now turn to discuss the ethical implications of this work. Worldly-Wise relies on leveraging cross-lingual knowledge resources for question answering. While this approach yields enormous benefits, care must be taken to evaluate appropriateness of the source of knowledge depending on the language. What may be considered as conventional wisdom in one culture or language may not be true for another. An example of how this manifests in our dataset is shown in Fig. 5. The knowledge graph conveys conventional wisdom in English that 'A dog is man's best friend', and therefore the expected answer to this question is 'Dog'. However, in regions where Hindi is spoken, the answer could equally be expected to be 'Cow' that appears in the image. This example is quite informative, and if such an instance can occur in the extreme, it could lead to fairness issues. This highlights the fundamental tradeoff involved in training such a cross-lingual system on knowledge generated in another language. Governance of such a system is therefore essential to ensure cultural appropriateness and fairness in different contexts.

[3] https://kaldi-asr.org/models/m1

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Zachary Ives, and et al. 2007. DBpedia: A nucleus for a web of open data. In *Proc. 6th International Semantic Web Conference*. Springer.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR*, abs/1506.02075.

Antoine Bordes, Nicolas Usunier, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *In Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *Proc. 25th AAAI Conference on Artificial Intelligence*, pages 301–306.

Tim Dettmers, Pascal Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2D knowledge graph embeddings. In *Proc. AAAI*, pages 1811–1818.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Patrick Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *Knowledge Distillation and Data Mining*, pages 601–610.

Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. 2013. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In *Proc. 22nd Internat. Conf. World Wide Web (WWW)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy.

Graeme S. Halford, William H. Wilson, and Steven Phillips. 2010. Review relational knowledge: the foundation of higher cognition.

David Harwath, Galen Chuang, and James Glass. 2018. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973, Calgary, AB.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *ACM International Conference on Web Search and Data Mining*.

Herman Kamper, Aristotelis Anastassiou, and Karen Livescu. 2019. Semantic query-by-example speech search using visual grounding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7120–7124. IEEE.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science*, 8693:740–755.

Hugo Liu and Push Singh. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22:211–226.

Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. 2017. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1211–1220, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems 31*, pages 2654–2665.

Medhini Narasimhan and Alexander G. Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *European Conference on Computer Vision (ECCV)*, pages 460–477.

M. Nickel, V. Tresp, and H.P. Kriegel. 2011. A three-way model for collective learningon multi-relational data. In *Proc. International Conference on Machine Learning*.

Kiran Ramnath and Mark Hasegawa-Johnson. 2020. Seeing is knowing! fact-based visual question answering using knowledge graph embeddings.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. 2019. KVQA: Knowledge-aware visual question answering. In *Proc. AAAI Conference on Artificial Intelligence*.

Leda Sarı, Samuel Thomas, and Mark Hasegawa-Johnson. 2020. Training spoken language understanding systems with non-parallel speech and text. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8109–8113. IEEE.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

R. Socher, D. Chen, C.D. Manning, and A.Y. Ng. 2013. Reasoning with neural tensor networks for knowledgebase completion. In *Advances in Neural Information Processing Systems*, page 926–934. Curran Associates, Inc.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *IN PROC. OF WWW '07*, pages 697–706. ACM.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *Proc. International Conference on Learning Representations*, pages 1–18.

Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. Webchild: harvesting and organizing commonsense knowledge from the web. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM*, pages 523–532, New York, NY, USA.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA. PMLR.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427.

William A. Woods. 1975. Motivation and overview of SPEECHLIS: An experimental prototype for speech understanding research. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 23(1):2–10.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–212.

Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. In *BMVC*.

ZFTurbo. 2018. Keras-retinanet for open images challenge 2018. https://github.com/ZFTurbo/Keras-RetinaNet-for-Open-Images-Challenge-2018.git.

Ted Zhang, Dengxin Dai, Tinne Tuytelaars, Marie-Francine Moens, and Luc Van Gool. 2017. Speech-based visual question answering. *arXiv preprint arXiv:1705.00464*.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452 – 1464.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1097–1103. International Joint Conferences on Artificial Intelligence Organization. Main track.