INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Towards Multi-Lingual Audio Question Answering

Swarup Ranjan Behera, Pailla Balakrishna Reddy, Achyut Mani Tripathi, Megavath Bharadwaj Rathod, Tejesh Karavadi

Reliance Jio - Artificial Intelligence Centre of Excellence (AICoE), Hyderabad, India

{swarup.behera, balakrishna.pailla, achyut.tripathi, megavath.rathod, tejesh.karavadi}@ril.com

Abstract

Audio Question Answering (AQA) is a multi-modal translation task where a system analyzes an audio signal and a natural language question to generate a desirable natural language answer. AQA has been primarily studied through the lens of the English language. However, addressing AQA in other languages, in the same manner, would require a considerable amount of resources. This paper proposes scalable solutions to multi-lingual audio question answering on both data and modeling fronts. We propose mClothoAQA, a translation-based multi-lingual AQA dataset in eight languages. The dataset consists of 1991 audio files and nearly 0.3 million question-answer pairs. Finally, we introduce a multi-lingual AQA model and demonstrate its strong performance in eight languages. The dataset and code can be accessed at https://github.com/swarupbehera/mAQA.

Index Terms: audio question answering, multi-lingual audio question answering, cross-modal task

1. Introduction

Audio Question Answering (AQA) is a multi-modal translation task where a system analyzes an audio signal and a natural language question to generate a desirable natural language answer [1]. Similar to visual question answering being important for enabling accessibility for visually impaired individuals [2, 3], AQA tasks have the potential to assist hearing-impaired individuals [4]. For example, a hearing-impaired person can utilize AQA on her wearable device to ask for and receive information about the surrounding acoustic environment in a question-answer style using natural language.

AQA is a core task in multi-modal research encompassing audio and language modeling. However, datasets for AQA task are only available in English. We present an example of AQA data in Table 1. AQA research has seen only three datasets, ClothoAQA [1], CLEAR [5], and DAQA [4]. ClothoAQA is a crowdsourced audio question-answering dataset generated from the audio files of the Clotho dataset [6], which contains day-to-day sounds occurring in the environment, such as water, nature, city, etc. In contrast, CLEAR and DAQA datasets are generated programmatically. Though CLEAR and DAQA have significantly more question-answer pairs than ClothoAQA, they lack diversity and challenges that are present in real-world data. We present a comparative overview of these three datasets in Table 2.

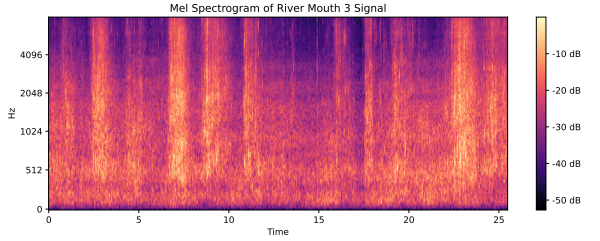
Because English-based AQA has garnered all the attention so far, AQA's promise applies exclusively to a privileged subset of human populations. In this paper, we take a step towards addressing this problem. We present a machine-translated multi-lingual audio question-answering dataset, mClothoAQA. mClothoAQA is built by generating multi-lingual question-answer pairs from the recently-proposed ClothoAQA dataset [1]. We also propose a multi-lingual AQA model to demonstrate the usage of our mClothoAQA dataset. In summary, our main contributions are

- We present a multi-lingual audio question answering dataset, mClothoAQA, in eight diverse languages: English (en), French (fr), Hindi (hi), German (de), Spanish (es), Italian (it), Dutch (nl), and Portuguese (pt). The mClothoAQA dataset comprises 1991 audio files and 35838 question-answer pairs for each language.
- We propose a multi-lingual AQA model and design two baseline experiments to show the usage of mClothoAQA:
  - A multi-modal multi-lingual binary classifier for 'yes' or 'no' type questions.
  - A multi-modal multi-lingual multi-class classifier for the other single-word answers.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 introduces the mClothoAQA dataset and presents its details. Section 4 presents the proposed multi-lingual AQA baseline model. Section 5 reports and analyzes the results of the experiments. Finally, Section 6 concludes the paper.

Table 1: *Sample audio question answering data from ClothoAQA [1] dataset containing the ⟨audio, question, answer⟩ triplets.*

Audio: river_mouth3.wav



Question	Answer
Are there waves?	Yes
How many times does the water splash?	Eleven
Is the area dry?	No
Is the sound falling water?	Yes
Is water flowing?	Yes
What is flowing?	Water

10.21437/Interspeech.2023-1816

Table 2: *Comparative overview of AQA datasets. Number of audios (#Audios). Audio duration in hours (Dur(h)). Maximum duration of audios in seconds (MD(s)). Average duration of audio in seconds (AD(s)). Question-answer source (QAS) indicates whether questions and answers are crowdsourced (C) or generated programmatically (P). Language (L) of the question and answer: English (E). Number of unique questions (#Ques). Number of unique answers (#Ans). Number of audio files in the training (Train), validation (Val), and test (Test) subsets.*

Dataset	#Audios	Dur(h)	MD(s)	AD(s)	QAS	L	#Ques	#Ans	Train	Val	Test
ClothoAQA [1]	1991	12.44	30	21	C	E	9153	830	1174	344	473
CLEAR [5]	50000	3.12	0.4	0.25	P	E	130957	47	35000	7500	7500
DAQA [4]	100000	2244.4	178.2	80.8	P	E	599294	36	80000	10000	10000

2. Related Work

Question Answering (QA) refers to the task of providing natural language answers to questions posed in natural language. There is a large body of work in QA depending on the type of additional input on which the question is targeted: (i) Textual Question Answering (TQA) [7, 8, 9] when a piece of text is used as an additional input, (ii) Visual Question Answering (VQA) [10, 11, 12, 13] when an image is used as an additional input, (iii) Audio Question Answering (AQA) [1, 5, 4] when an audio signal is used as an additional input, and (iv) Video Question Answering (VideoQA) [14, 15, 16, 17, 18, 19, 20] when a video is used as an additional input.

AQA is a relatively new area of research. To the best of the authors' knowledge, only three works have been published in AQA research so far: ClothoAQA [1], CLEAR [5], and DAQA [4]. In terms of data, the CLEAR dataset is similar to the DAQA dataset. These are both synthetic datasets of audio sequences, questions, and answers. CLEAR and DAQA datasets are generated programmatically, while ClothoAQA is a crowdsourced dataset generated from the audio files of Clotho dataset [6], an audio captioning dataset. Although CLEAR and DAQA have significantly more question-answer pairs than ClothoAQA, they lack diversity and challenges that are present in real-world data. This is why we built our multi-lingual mClothoAQA dataset from the ClothoAQA dataset. Refer to Table 2 for more details on these datasets. In terms of modeling, all of the previous models on AQA [1, 4] are built for English. Further, it treats AQA as a classification task over a pre-defined set of the top (English) answers. In this paper, we are introducing a multi-lingual AQA dataset and a multi-lingual AQA model for the first time.

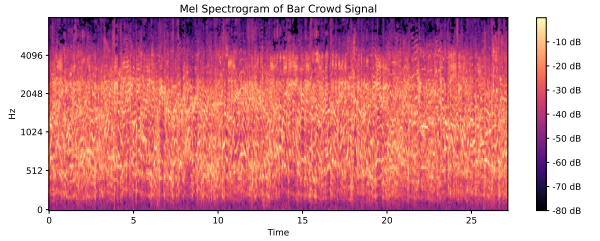
3. mClothoAQA Dataset

Like most machine learning tasks, obtaining high-quality labeled data is the main bottleneck for multi-lingual AQA. To address this, we have built a multi-lingual AQA dataset called mClothoAQA. We translated the questions and answers from the ClothoAQA [1] dataset into seven other languages using Google's machine translation API [1]. We also tested a few other open-source machine translation tools, but their translation quality did not match that of Google's. We evaluated the translations using metrics such as BLEU [21], ROUGE [22], and METEOR [23]. For this work, we selected seven diverse languages: French (fr), Hindi (hi), German (de), Spanish (es), Italian (it), Dutch (nl), and Portuguese (pt). After the machine translation, the question-answer pairs were verified and adjusted by humans. Please refer to Table 3 for examples of mClothoAQA.

[1] https://cloud.google.com/translate

Table 3: *mClothoAQA dataset in eight languages. From top to bottom: English (en), French (fr), Hindi (hi), German (de), Spanish (es), Italian (it), Dutch (nl), and Portuguese (pt).*

Audio: bar crowd.wav



Lang	Question	Answer
en	Are many people talking?	Yes
fr	Beaucoup de gens parlent-ils?	Oui
hi	क्या कई लोग बात कर रहे हैं?	हाँ
de	reden viele Leute?	Ja
es	¿Están hablando mucha gente?	Sí
it	Molte persone parlano?	SÌ
nl	praten veel mensen?	Ja
pt	Muitas pessoas estão falando?	sim

The English subset of the mClothoAQA dataset (mClothoAQA-en) is similar to the ClothoAQA dataset. The mClothoAQA-en subset consists of 1991 audio files that capture day-to-day sounds occurring in the environment, such as water, nature, birds, noise, rain, city, wind, and more. Each audio file is associated with six questions. In total, mClothoAQA-en contains 9153 unique questions. These questions can be categorized into two types: (i) 'yes' or 'no' answer type questions (four questions), and (ii) single-word answer type questions (two questions). Each question has three answers from three annotators. Therefore, each of the 1991 audio files has 18 question-answer pairs. The mClothoAQA-en subset includes a total of 830 unique words as answers.

Non-overlapping training, validation, and testing splits are created from mClothoAQA-en in the ratio of 60%-20%-20% using multilabel-stratification [24]. The training, validation, and test splits of mClothoAQA-en consist of 1174, 344, and 473 audio files, respectively, and include 830, 512, and 801 unique answers. In Table 2, we provide a comparative overview of various statistics for the mClothoAQA-en (or ClothoAQA) dataset.

We present the breakdown of question types in the training set of mClothoAQA-en in Figure. 1. It illustrates the distribution of the first four words for all the questions in the training set of mClothoAQA-en, showcasing diverse question types and

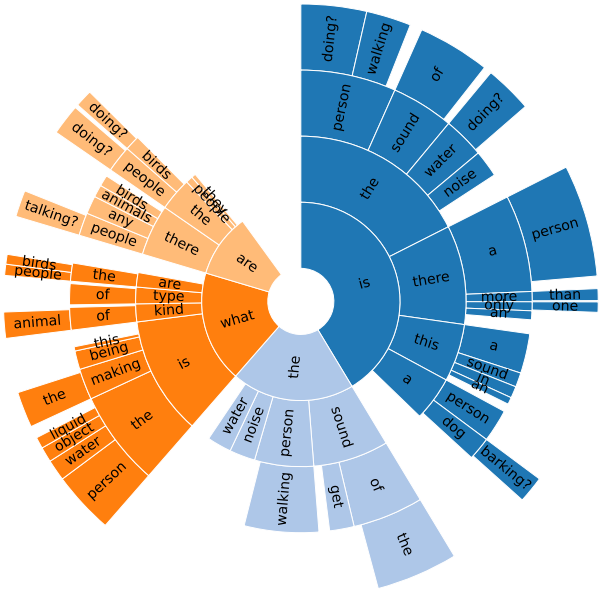


Figure 1: *Distribution of the first four words for all questions in the training set of mClothoAQA-en dataset. The innermost ring represents the first word, and radiating rings represent subsequent words. Arc lengths are proportional to the number of questions containing the word. Words with a frequency less than 35 are omitted for clarity.*

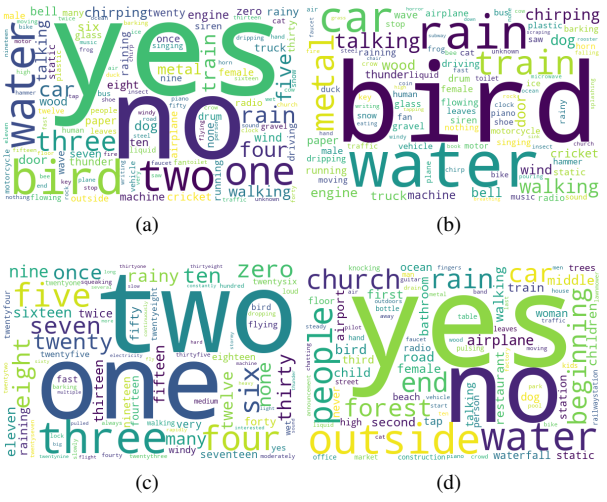


Figure 2: *Word clouds of the answers for different question types in the training set of mClothoAQA-en. (a) all the questions, (b) 'what' type questions, (c) 'how many' type questions, and (d) questions other than 'what' and 'how many' type such as 'is/are/was/were/do/does/did/the'.*

a high degree of linguistic variations. Additionally, Figure. 2 showcases word clouds of answers for selected question types in the training set of mClothoAQA-en, further highlighting diverse answers for each question type.

Similar analyses can be conducted for the other language subsets (e.g., mClothoAQA-hi, the Hindi subset of mClothoAQA) of the mClothoAQA dataset.

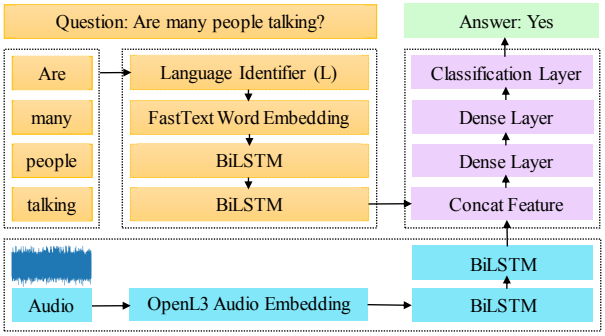


Figure 3: *Baseline multi-lingual AQA model architecture.*

4. Modeling

All of the previous work on AQA has been built for English. Further, it treats AQA as a classification task over a pre-defined set of the top (English) answers. In this section, we present a multi-lingual AQA baseline model.

Our Bidirectional Long Short-Term Memory (BiLSTM) [25] based model architecture for multi-lingual AQA is shown in Figure 3. The inputs to the model are the audio signal and the question (in a particular language) related to the audio. The mClothoAQA dataset has a relatively small amount of audio files and texts (questions) to learn a good representation. We take advantage of existing pre-trained audio embeddings (OpenL3 [26]) and text embeddings (fastText [27]) to extract audio and text features, respectively. Both the audio and text sides extract their respective audio features and text features and generate fixed-size representations. These representations are concatenated and passed through fully connected layers for classification. The model's output is either 'yes' or 'no' in the case of binary classification or one of 828 single-word answers in the case of multi-class classification.

On the audio side, we first extract $\mathbf{X} \in \mathbb{R}^{T \times 128}$ Mel spectrogram from the input audio signal with 128 mel bands and T time frames. The Mel spectrogram is then fed to OpenL3 [26] environmental model to compute $\mathbf{X}_{emb} \in \mathbb{R}^{T' \times 512}$ deep audio embeddings, where T' is the number of output time frames from the OpenL3 model and 512 is the audio embedding size. These embeddings are extracted using a hop size of 0.1 seconds. In order to understand temporal correlations and transform them into a fixed-size representation, the audio embeddings are then processed through a succession of BiLSTM layers, $BiLSTM_n$ with $n = 1, 2$. The bidirectional LSTM is given by

$$X_n = BiLSTM_n(X_{n-1}) \tag{1}$$

where $X_0 = X_{emb}$. If h is the number of hidden units in the BiLSTM, then $\mathbf{X}_n \in \mathbb{R}^{T' \times 2h}$. We then choose as output $\mathbf{x}_n \in \mathbb{R}^{2h}$, the final time step of the last BiLSTM layer to represent the fixed-size audio embedding.

On the text side, a Language Identifier (LangId = L) module specifies the language of the input question. The LangId is used to select the pre-trained text embeddings for that specific language. To encode the input question into word embeddings, we utilize FastText [27] pre-trained word vectors. These word vectors for 157 languages can be accessed here [2]. If the input question Q has K words, the word embeddings using FastText

[2] https://tinyurl.com/w658k9ah

are denoted as $\mathbf{Q}emb \in \mathbb{R}^{K \times 300}$, where 300 represents the dimensionality of the word embeddings. The word embeddings are then processed through a series of bidirectional LSTM layers. If $h^{'}$ is the size of the hidden units in the BiLSTM, we again choose as output $\mathbf{q}_n \in \mathbb{R}^{2h^{'}}$, the final time step of the last BiLSTM layer to represent the fixed size word embedding for the input question.

For both the audio and text side, we use a hidden size of 128 for the BiLSTM layers with a dropout of 0.2 for the binary classifier and a hidden size of 512 for the multi-class single-word answers classifier. The audio and text side outputs are concatenated and passed through a series of fully connected layers $Dense_k$ with $k = 1, 2$ with ReLU non-linearity. The fully connected layers combine the learned features of both the audio and the textual question. This is given by

$$D_n = Dense_k(D_{k-1}) \tag{2}$$

where $\mathbf{D}_0 = concat\,[\mathbf{x}_n, \mathbf{q}_n]$, and *concat* represents concatenation operator. The binary classifier's two dense layers have 256 and 128 neurons each. However, the multi-class classifier's two dense layers have 1024 neurons each, increasing the model's capacity to capture finer details and classify among 828 answer classes.

The last dense layer's output is then passed to a classification layer, which is another dense layer with the number of neurons equal to the number of distinct answer classes. The output of the classification layer is $\hat{y} \in [0, 1]^C$, where *C* represents the number of answer classes in the dataset. For the 'yes' or 'no' binary classifier, the classification layer is a simple logistic regressor with one neuron. In the multi-class classifier, the classification layer comprises 828 neurons, each representing a distinct single-word answer. The softmax activation function is then applied after the classification layer.

5. Experimental Results and Analysis

We trained and evaluated the multi-lingual AQA models separately on each of the eight languages of mClothoAQA dataset.

In order to create the splits for the binary classifier, we select the 'yes' or 'no' questions from the associated data splits. 1174, 344, and 473 audio files with twelve 'yes' or 'no' question-answer pairings per file are used for training, validation, and testing, respectively. We train and evaluate the binary classifier on the following three subsets of data to examine how well it performs when different annotators provide contradictory answers to the same question.

- **Unfiltered:** All the question-answer pairs are considered, even if they have contradicting answers.
- **Unanimous:** Only those question-answer pairs are considered where all three annotators have responded unanimously.
- **Majority:** For each question, the answer provided by at least two of the three annotators is considered.

Similarly, for the multi-class classification task, we select the single-word answers from the respective data splits, resulting in the same number of audio files as the original splits, with each file having six question-answer pairs.

All the models are trained with the Adam optimizer with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The models are trained for 100 epochs with cross-entropy loss, and the model with the best validation score is used for testing. We use 'accuracy' as the evaluation metric for the binary classification task. For the multi-class classification task, we use 'top-1', 'top-5', and 'top-10' accuracy. The 'top-1' accuracy is conventional, i.e., the model's answer must match the expected answer. The 'top-5' accuracy means any of the model's five highest probability answers must match the expected answer.

Table 4: *Accuracies (%) of binary 'yes' or 'no' classifier on multi-lingual AQA data.*

Languages	Unfiltered	Unanimous	Majority
English (en)	62.05	**71.50**	63.00
French (fr)	60.28	67.87	62.94
Hindi (hi)	**62.61**	70.64	**65.09**
German (de)	61.20	68.95	62.31
Spanish (es)	61.36	69.43	61.36
Italian (it)	60.97	67.14	62.47
Dutch (nl)	62.35	69.67	64.21
Portuguese (pt)	62.57	69.43	64.21

Table 5: *Accuracies (%) of single-word answer multi-class classifier on multi-lingual AQA data.*

Languages	Top-1 Acc	Top-5 Acc	Top-10 Acc
English (en)	**53.73**	**91.68**	**98.57**
French (fr)	51.96	90.24	95.93
Hindi (hi)	53.31	91.54	97.41
German (de)	51.29	89.50	96.52
Spanish (es)	52.07	90.19	97.09
Italian (it)	51.62	89.69	95.70
Dutch (nl)	53.09	91.05	95.38
Portuguese (pt)	52.48	90.70	96.96

Table 4 summarises the results of our binary classification experiments on the mClothoAQA dataset for each of the eight languages. It is evident from the results that the model performs better when the answers are unanimous, indicating the intelligible presence of the answer in the audio.

The results of our single-word multi-lingual multi-class classifier are presented in Table 5. We also use top-5 and top-10 accuracy metrics to evaluate the classifier's performance, considering the high number of unique answer classes (828). The results indicate that the model is beginning to capture the relationships between the multi-modal data.

This is the first work on multi-lingual AQA; therefore, conducting a comparative analysis is out of scope. Our future goals to enhance multi-lingual AQA classification accuracy are as follows: (i) developing a cross-lingual, unified, extensible, open-ended, and end-to-end transformer-based mAQA model, and (ii) creating a scalable translation-based framework for generating high-quality mAQA data based on audio captions.

6. Conclusions

We take initial steps towards multi-lingual AQA by proposing scalable solutions for data creation and modeling. Our contributions include the creation of a multi-lingual AQA dataset in eight diverse languages, aimed at driving progress in multi-lingual AQA modeling. Additionally, we establish a baseline multi-lingual AQA model that demonstrates good performance across the eight languages. We believe that the mClothoAQA dataset, the multi-lingual AQA model, and the experimental results presented in this work will serve as benchmarks for future research in multi-lingual audio question answering.

7. References

[1] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1140–1144.

[2] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 3608–3617.

[3] W. S. Lasecki, P. Thiha, Y. Zhong, E. L. Brady, and J. P. Bigham, "Answering visual questions with conversational crowd assistants," in *15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. ACM, 2013, pp. 18:1–18:8.

[4] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 2283–2294, 2020.

[5] J. Abdelnour, G. Salvi, and J. Rouat, "Clear: A dataset for compositional language and elementary acoustic reasoning," 2019. [Online]. Available: https://dx.doi.org/10.21227/7x26-a025

[6] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. The Association for Computational Linguistics, 2016, pp. 2383–2392.

[8] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," in *56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2018, pp. 784–789.

[9] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," in *2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL*. Association for Computational Linguistics, 2017, pp. 191–200.

[10] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "VQA: visual question answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[11] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3–20, 2017.

[12] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," in *Annual Conference on Neural Information Processing Systems*, 2015, pp. 2296–2304.

[13] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," *International Journal of Computer Vision*, vol. 127, no. 4, pp. 398–414, 2019.

[14] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "TVQA: localized, compositional video question answering," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2018, pp. 1369–1379.

[15] K. Kim, M. Heo, S. Choi, and B. Zhang, "Deepstory: Video story QA by deep embedded memory networks," in *26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 2016–2022.

[16] D. Patel, R. Parikh, and Y. Shastri, "Recent advances in video question answering: A review of datasets and methods," *CoRR*, vol. abs/2101.05954, 2021.

[17] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *33rd AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2019, pp. 9127–9134.

[18] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2016, pp. 4631–4640.

[19] K. Zeng, T. Chen, C. Chuang, Y. Liao, J. C. Niebles, and M. Sun, "Leveraging video descriptions to learn video question answering," in *31st AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2017, pp. 4334–4340.

[20] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 409–421, 2017.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *40th Annual Meeting on Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2002, pp. 311–318.

[22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, 2004, pp. 74–81.

[23] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2007, pp. 228–231.

[24] K. Sechidis, G. Tsoumakas, and I. P. Vlahavas, "On the stratification of multi-label data," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, vol. 6913. Springer, 2011, pp. 145–158.

[25] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673 – 2681, 1997.

[26] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[27] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," *CoRR*, vol. abs/1712.09405, 2017.