Multi-Level Multimodal Transformer Network for Multimodal Recipe Comprehension

Ao Liu[1], Shuai Yuan[1], Chenbin Zhang[1,3], Congjian Luo[1,2], Yaqing Liao[1], Kun Bai[3], Zenglin Xu[4,2,1,*]
[1] SMILE Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2] Artificial Intelligence Center, Peng Cheng Lab, Shenzhen, China
[3]Cloud and Smart Industries Group, Tencent, China
[4] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
{zeitmond,shuaiyuan0209,aleczhang13,im.congjian,yaqingliao1997}@gmail.com,kunbai@tencent.com,zenglin@gmail.com

ABSTRACT

Multimodal Machine Comprehension (M^{3}C) has been a challenging task that requires understanding both language and vision, as well as their integration and interaction. For example, the RecipeQA challenge, which provides several M^{3}C tasks, requires deep neural models to understand textual instructions, images of different steps, as well as the logic orders of food cooking. To address this challenge, we propose a Multi-Level Multi-Modal Transformer (MLMM-Trans) framework to integrate and understand multiple textual instructions and multiple images. Our model can conduct intensive attention mechanism at multiple levels of objects (*e.g.*, step level and passage-image level) for sequences of different modalities. Experiments have shown that our model can achieve the state-of-the-art results on the three multimodal tasks of RecipeQA.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

multimodal machine reading comprehension, multimodal recipe comprehension, question answering

ACM Reference Format:
Ao Liu, Shuai Yuan, Chenbin Zhang, Congjian Luo, Yaqing Liao, Kun Bai and Zenglin Xu. 2020. Multi-Level Multimodal Transformer Network for Multimodal Recipe Comprehension. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401247

1 INTRODUCTION

Machine Reading Comprehension (MRC) is an important research topic in the cross-discipline area of information retrieval, natural language processing and machine learning. Recently, Multimodal Machine Comprehension (M^{3}C) [8], as a multimodal extension of the traditional MRC task, has emerged with multiple datasets [7, 8, 21]. It differs from standard MRC and Visual Question Answering(VQA) [1] in that the context or questions may be multimodal, *e.g.* containing both text and images.

M^{3}C brings new challenges: 1) compared with standard MC tasks(e.g., [14, 16, 23]), it requires comprehending both text and images and capturing the text-to-image semantic correlations; 2) compared with VQA tasks(*e.g.* [10]), it requires learning rich semantics (*e.g.* the orders of instructions) from abundant passages and multiple images, while in VQA the context is a single picture and the question is just a sentence.

The RecipeQA challenge [21] is a recent M^{3}C benchmark, which focuses on multimodal recipe comprehension and provides four

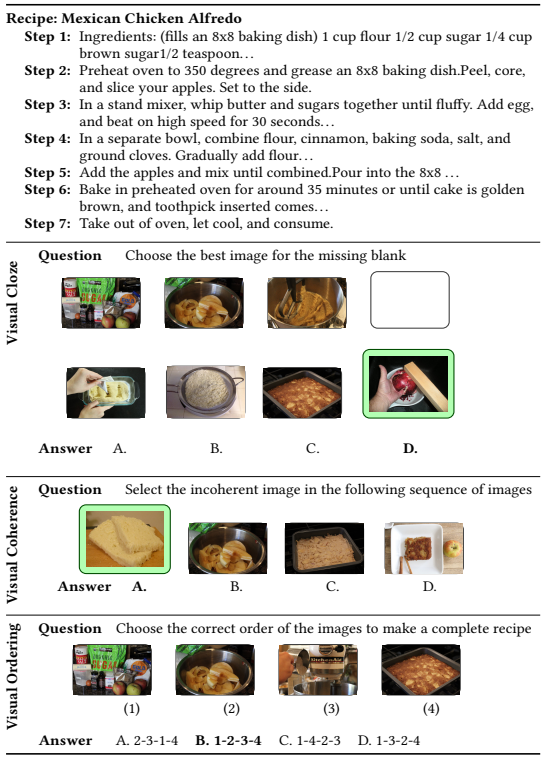

Figure 1: A sample of *visual cloze, visual coherence* and *visual ordering*, which has common context comprised of step descriptions. The correct answers are in green frames or in bold.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401247

*Corresponding author.

different multi-choice tasks, three of which show a special form of M^{3}C: textual context and visual questions. These tasks require different reasoning skills. As shown in Figure 1, the *visual cloze* task tests a skill to infer a missing image from a coherent sequence of images. The *visual coherence* task tests the ability to identify an incoherent image from the ordered set of images. The *visual ordering* task requires the capability to sort a set of jumbled images. Similar task forms can also be seen in COMICS [7] which also contains visual cloze and coherence tasks, but it focuses on comic book panels along with captions, while RecipeQA provides real-world pictures taken by random users from the Internet, strictly step-wise instructions, and multiple tasks which have common forms. These characteristics make RecipeQA potentially more practical to real-world applications. Besides, Multimodal Recipe Comprehension is very close to Cross-modal Recipe Retrieval [3, 17], which requires retrieving images given recipe text and vice versa. The key problem of Cross-modal Recipe Retrieval is how to match corresponding recipes and images, which is also significant for RecipeQA.

To the best of our knowledge, there have not been state-of-the-art models addressing the M^{3}C task where multiple sentences and multiple images are to be matched together and the coherence of images is necessary to be learned. Previous vision+language models [18, 22] often aim at the case where a single sequence of words and local image regions (like "visual words") are matched, which limit themselves to a lower semantic level, not suitable for our case. An earlier work on RecipeQA [12] has achieved the state-of-the-art result on the *textual cloze* task in a uni-modal (text only) form. However, the other three *visual-* tasks still require further exploration.

To address these problems in these three *visual*-tasks of RecipeQA in a unified manner, we propose a Multi-Level Multi-Modal Transformer (MLMM-Trans). Our model considers multiple levels of attention: step-level and passage-image level. We use multi-head self-attention to learn the dependencies between the steps in the passage, to which we refer as the step-level attention, and conduct passage-image attention to learn the correlations between each step in the context and each image in the question, based on our modified Transformers [20] block. Existing work on multimodal Transformer-like architectures [11, 13, 18] are usually based on a pretrained language model BERT [4] and depend a lot on large amounts of data. Also, they usually concatenate text and images together and adopt a self-attention scheme upon them to discover implicit alignments between language and vision, which can be useful when a sequence of words and image regions are matched but has not been proved effective when a sequence of sentences(steps of a recipe) and a sequence of images are to be matched. Moreover, the steps do not have as explicit contextual relationships as words do and could not exploit the pretrained contextual language models. VideoBERT [18] chose to merge all the sentences into a long sequence, but it's not plausible for RecipeQA since some recipes in RecipeQA tend to be very very long and the lengths vary to an extreme extent(from 3 steps to 25 steps). Our model, however, does not mix the text and images but put them into an interactive attention module and refine the fused representation for multiple times. The intensive attention computation guarantee the implicit co-alignment of multimodal data. Furthermore, experiments demonstrate that our model can achieve state-of-the-art results on all *visual*-tasks of RecipeQA, which can also serve as a strong baseline for the RecipeQA challenge. Note that although the proposed method was evaluated only on the RecipeQA challenge (the only available dataset with rich multimodality semantics), it can be insightful to other retrieval tasks or applications with consideration to orders and coherence in multimodalities, *e.g.* retrieving a series of images given text.

2 METHOD

In this section, we formalize the problem and explain our model. We treat the RecipeQA tasks as a classification problem. Given a passage $D = \{S_1, S_2, ...S_N\}$ composed of N coherent steps, where a step $S_i = \{w_1^i, w_2^i, ...w_{L_S}^i\}$ has L_S tokens, the model needs to choose from K candidate answers $\{a_1, a_2, ...a_K\}$. Each answer $a_k = \{I_1^k, I_2^k, ...I_m^k\}$ is constructed by combining the original question and an answer k in the tasks and composed of m images. We aim to capture the correlations between the sequence of steps D and sequences of images a_k-s. A most correlated image sequence is chosen as the final answer, *i.e.*, $\hat{a} = \text{argmax}_k P(a_k, D)$. The overview of our model is shown in Figure 2. We perform different pre-processing settings for different tasks to transform each question-answer pair into a sequence of images, and the number m of images in an answer could vary for different tasks. For *visual cloze*, we fill the candidate answers into the question to form four sequences of images with number $m = 4$. For *visual coherence*, we eliminate each of the four candidate images from the question sequence to construct four sequences of images with number $m = 3$. For *visual ordering* ,we construct four sequences of images with the four candidate orders, making $m = 4$.

2.1 Encoding Multi-modalities

Each image is fed into a pre-trained ResNet-50[5] to obtain a d_V-dimensional vector. Then it's projected to the hidden dimension d with a linear transformation.

$$\mathbf{h}^{I_j} = \mathbf{W}^I(\text{CNN}(I_j)), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times d_V}$ is a learned parameter. For each candidate image sequence a_k, we concatenate all the images to obtain the visual matrix $\mathbf{H}_k^V \in \mathbb{R}^{m \times d}$:

$$\mathbf{H}_k^V = [\mathbf{h}_k^{I_1}, \mathbf{h}_k^{I_2}, \cdots, \mathbf{h}_k^{I_m}]. \tag{2}$$

Each token in the steps is first embedded as a d_T-dimensional pre-trained vector. We further encode each step S_i with a bi-directional LSTM[6] followed by max-pooling.

$$\mathbf{h}^{S_i} = \text{MaxPooling}(\text{Bi-LSTM}(S_i)), \tag{3}$$

where $\mathbf{h}^{S_i} \in \mathbb{R}^d$ and d is the hidden size of the Bi-LSTM.

Thus we obtain the text matrix consisting of context-dependent step representations as $\mathbf{H}^D = [\mathbf{h}^{S_1}, \mathbf{h}^{S_2}, \cdots, \mathbf{h}^{S_N}]$, where $\mathbf{H}^D \in \mathbb{R}^{N \times d}$ is the concatenation of all step-level representations.

2.2 Step-level Attention

We modify the Transformers architecture [20] to build our attention layers. Given packages of a set of d-dimensional queries, keys, and values, denoted as $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$, we first project them H times to d/H dimensions with different linear transformations. Then the scaled

dot-product attention is computed with a scaling factor of $1/\sqrt{d/H}$. The outputs are then concatenated together followed by a linear projection, as follows:

$$\hat{\mathbf{Q}}^i = \mathbf{Q}\mathbf{W}_Q^i,\ \hat{\mathbf{K}}^i = \mathbf{K}\mathbf{W}_K^i,\ \hat{\mathbf{V}}^i = \mathbf{V}\mathbf{W}_V^i,$$

$$\hat{\mathbf{M}}^i = \text{softmax}(\frac{\hat{\mathbf{Q}}^i\hat{\mathbf{K}}^{iT}}{\sqrt{d/H}})\hat{\mathbf{V}}^i,$$

$$\text{MH}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\hat{\mathbf{M}}^1;\hat{\mathbf{M}}^2;\cdots;\hat{\mathbf{M}}^H]\mathbf{W},$$

where $\mathbf{W}_Q^i, \mathbf{W}_K^i, \mathbf{W}_V^i \in \mathbb{R}^{d\times d/H}$ are the parameter matrices for the i-th head and $\mathbf{W} \in \mathbb{R}^{d\times d}$. To learn the dependency relationships among steps in the context, we utilize a self-attention mechanism on top of the step representations $\mathbf{H}^D$. The passage encoder is composed of L stacked identical attention layers. For each layer l, we update $\mathbf{Q}^l = \mathbf{K}^l = \mathbf{V}^l = \mathbf{E}^l = F(\mathbf{E}^{l-1})$ and set $\mathbf{E}^0 = \mathbf{H}^D$.

$$\text{G}_s(\mathbf{E}^l) = \text{MH}(\mathbf{E}^l,\mathbf{E}^l,\mathbf{E}^l) + \mathbf{E}^l, \tag{4}$$

$$\text{F}(\mathbf{E}^l) = \text{FF}(\text{G}_s(\mathbf{E}^l)) + \text{G}_s(\mathbf{E}^l), \tag{5}$$

$$\text{FF}(\mathbf{x}) = \text{ReLU}(\mathbf{x}\mathbf{W}_1^l + \mathbf{b}_1^l)\mathbf{W}_2^l + \mathbf{b}_2^l, \tag{6}$$

where $\text{FF}(\cdot)$ denotes a positional feed-forward network. We then apply layer normalization [2] on the final-layer output $\mathbf{E}^L$, i.e.,

$$\hat{\mathbf{E}} = \text{LayerNorm}(\mathbf{E}^L). \tag{7}$$

2.3 Passage-Image Level Attention

After obtaining the interdependent representations of the steps in the context paragraph, we employ a passage-image level attention mechanism, where we aim to learn the relationship of the sequence of steps and sequences of images. The same multi-head attention block is utilized but with different queries. Still, we have L stacked attention layers. For each layer l, we set the input $\mathbf{Q}^l = \mathbf{H}_k^V$ permanently but update $\mathbf{K}^l = \mathbf{V}^l = \mathbf{U}^l = \text{F}(\mathbf{U}^{l-1})$ to model the image-to-text attention and $\mathbf{U}^0 = \mathbf{H}^D$. In this way, information in textual context can be fused into image representations. The computation of layer l is as follows:

$$\text{G}_t(\mathbf{U}^l) = \text{MH}(\mathbf{Q}^l,\mathbf{U}^l,\mathbf{U}^l) + \mathbf{Q}^l, \tag{8}$$

$$\text{F}(\mathbf{U}^l) = \text{FF}(\text{G}_t(\mathbf{U}^l)) + \text{G}_t(\mathbf{U}^l). \tag{9}$$

Similar to Equation (7), the final-layer output $\mathbf{U}^L \in \mathbb{R}^{m\times d}$ is transformed to $\hat{\mathbf{U}}$ via layer normalization. After the first layer of the above computation, we obtain a new visual representation $\mathbf{U}^1$ by weighted sum of the step vectors. The same computation repeats for several layers, where the new representation is refined multiple times, guided by the original image vectors $\mathbf{H}_k^V$.

2.4 Sequential Modeling

After obtaining the attended representation of image sequence a_k with respect to steps, we concatenate it with the original image representation to fully utilize the image information, i.e., $\mathbf{V}_k = [\mathbf{H}_k^V;\hat{\mathbf{U}}]$, where $\mathbf{V}_k \in \mathbb{R}^{m\times 2d}$ is the text-dependent image matrix for a candidate image sequence. Next, we model the coherence degree of this sequence with the context passage. Due to the effectiveness of LSTMs on modeling sequential data, we utilize a uni-directional

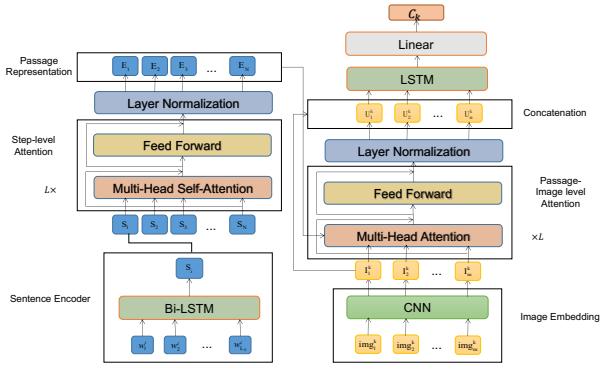

Figure 2: An overview of our proposed model. It builds a correlated representation with modified Transformer blocks that combine the representations for the context text and an answer sequence. The final score C_k for answer a_k is fed into a softmax classifier to infer the answer.

LSTM followed by max pooling and linear projection to obtain the coherence score for this candidate answer as follows

$$\mathbf{o}_k = \text{MaxPooling}(\text{LSTM}(\mathbf{V})), \tag{10}$$

$$C_k = \mathbf{o}_k\mathbf{W}^o + b^o, \tag{11}$$

where $\mathbf{W}^o \in \mathbb{R}^{d\times 1}$ and $b^o \in \mathbb{R}$ are parameters and the hidden dimension of the LSTM is d. C_k is the final output of the matching between the sequences of the passage and the candidate answer sequence a_k. Note that we make special modifications for the *visual coherence* task. Since the aim of *visual coherence* is to find the most incoherent image among the sequence of images, we measure the distance between each image and all the other images in the question with L_2-norm and add the average distance to the matching score as a regularizer.

2.5 Answer Prediction

Similarly, we obtain the matching scores for all the K candidate answer sequences. The correct answer is predicted through a softmax classifier:

$$P(a_k|D) = \frac{\exp C_k}{\sum_{j=1}^K \exp C_j}. \tag{12}$$

3 EXPERIMENTS

We perform evaluations on these three *visual-* tasks (*visual cloze, visual coherence, visual ordering*) of RecipeQA dataset. Each of them has about 7k training data, 1k validation data and 1k test data. We report our result on the publicly available test sets after validated on the validation sets. We do not submit our results to the official leaderboard due to some technical problems of the submission system.

3.1 Implementation Details

Pre-trained 300-dimensional GloVe [15] embedding is utilized for word embedding and the weights are frozen during training. The dimension d_V for visual encoder is 2048. We do not utilize BERT [4]

	Visual Cloze	*Visual Coherence*	*Visual Ordering*
HUMAN	77.60	81.60	64.00
Hasty Student	27.35	65.80	40.88
Impatient Reader	27.36	28.08	26.74
PRN (Single Task)	56.31	53.64	62.77
PRN (Multi Task)	46.45	40.58	62.67
MM-Trans	63.80	62.63	60.88
MLMM-Trans	**65.57**	**67.33**	**63.75**

Table 1: Accuracy on the test sets of the *visual*-tasks of RecipeQA.

to encode our step-level representations since we did not observe obvious increases on the performance, which may result from the noise in the original recipe text—the recipes are written in unlimited environments and contain a lot of out-of-vocabulary tokens. An Adamax [9] optimizer is used with learning rate 0.002 and the batch size is fixed to 20. The hidden size is set to $d = 512$ and for attention layers, we set $L = 6$ and $H = 8$ to keep consistent with the original Transformers [20] settings. We do not tune these hyperparameters after initialization, which indicates further potential improvement.

3.2 Experimental Results

We compare our model with the existing baselines including the leaderboard results of the RecipeQA Challenge. **Impatient Reader** [21] uses a uni-directional LSTM with 3 layers to encode the text and images and a hinge ranking loss for training. Small adjustments are made for different tasks. **Hasty Student**[21] is adapted from [19]. It only considers the similarities or dissimilarities between the images in the questions and candidate answers. **PRN**(Single Task / Multi Task) is the current champion on the RecipeQA leaderboard[*]. Besides, **HUMAN** accuracy is reported by the authors of RecipeQA.

To further demonstrate the effectiveness of our model, we also make an extra baseline model. For fair comparison, and due to the lack of strong baselines, we want to take previous multimodal Transformers architectures into account, like VideoBERT [18], ViLBERT [13], etc. However, these BERT-like models require large-scale data to support their effects. Also, they focus on different tasks and could not be directly applied to RecipeQA tasks. Therefore, we adopt a similar and easy-to-implement scheme, to concatenate the sentences and images together and feed the new sequence into a Transformer block. In other words, we replace our multi-level attention layers with self-attention to capture the relationship between steps and images. The other partial modules remain the same. We call this model MM-Trans for clarity.

As shown in Table 1, our approach outperformed all the existing models on all three tasks. We can also notice that the performance on the *visual ordering* task reaches very close to human accuracy. Compared with MM-Trans, our model achieves much better accuracy, which shows the superiority over simple self-attention.

4 CONCLUSION

We have presented a Multi-level Multimodal Transformer framework to address the multimodal comprehension task. By modeling interactions between different modalities at multiple levels, we have achieved state-of-the-art performance on three different tasks of RecipeQA.

[*]https://hucvl.github.io/recipeqa/

ACKNOWLEDGEMENT

This work was partially supported by the National Key Research and Development Program of China (No. 2018AAA0100204).

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*. 2425–2433.
[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
[3] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *SIGIR*. 35–44.
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[7] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *CVPR*. 7186–7195.
[8] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*. 4999–5007.
[9] Diederik P Kingma and Jimmy Ba. 2014. *arXiv preprint arXiv:1412.6980* (2014).
[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
[11] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).
[12] Ao Liu, Lizhen Qu, Junyu Lu, Chenbin Zhang, and Zenglin Xu. 2019. Machine Reading Comprehension: Matching and Orders. In *CIKM*. 2057–2060.
[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. 13–23.
[14] Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing Interpretive Spatio-Temporal Features for Multi-Turn Responses Selection. In *ACL'19*. 44–50.
[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.
[16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).
[17] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *CVPR*. 3020–3028.
[18] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *ICCV*. 7464–7473.
[19] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *CVPR*. 4631–4640.
[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[21] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In *EMNLP*. 1358–1368.
[22] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal Transformer with Multi-View Visual Representation for Image Captioning. *arXiv preprint arXiv:1905.07841* (2019).
[23] Chenbin Zhang, Congjian Luo, Junyu Lu, Ao Liu, Bing Bai, Kun Bai, and Zenglin Xu. 2020. Read, Attend, and Exclude: Multi-Choice Reading Comprehension by Mimicking Human Reasoning Process. In *SIGIR*.