View Meta-Reviews

Paper ID
9802

Paper Title
MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

Track Name
CVPR2022

META-REVIEWER #1

META-REVIEW QUESTIONS

3. Paper final decision
Accept

5. Decision summary
Accept: The paper received all accept recommendations. The area chairs agree with this recommendation. This decision has been confirmed by the AC panel.

7. Comments on decision
This paper received acceptance recommendations from all three reviewers after the rebuttal. Only R1 participated in the discussion. Overall, the experiments, the large-scale dataset and the multi-modal contrastive masked span training were appreciated by the reviewers. AC is therefore happy to accept this paper at CVPR. This decision was confirmed at the AC triplet. The authors are invited to incorporate the materials from the rebuttal (e.g., action recognition results) in their final version, as well as completing the promised experiments for R1. Since the paper claims a dataset contribution, according to CVPR rules, it is expected that the dataset will be made publicly available no later than the camera-ready deadline.

META-REVIEWER #2

META-REVIEWER #3

View Reviews

Paper ID
9802

Paper Title
MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

Track Name
CVPR2022

Reviewer #1

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
This paper represents a RESERVE model designed to learn a union representation from multi-modality input and benefit various downstream tasks. RESERVE is a transformer-based model with multi-modal input (video/text/audio). RESERVE is trained self-supervised and with a novel objective, namely contrastive span training. The authors claim that such objective function benefits cross-modality representation learning. The experiments show that REVERSE achieves the best performance on VCR and TVQA. Furthermore, Zero-shot experiments further demonstrate RESERVE received noticeable gain through span-style pertaining at scale and multi-modal representation.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.

This paper is well-written and clear enough to follow the core idea. Moreover, the authors did an excellent job distinguishing their work from previous works.
The experiments are solid, thorough and well-documented with various detailed settings, giving it high confidence to claim that REVERSE and contrastive span training can benefit future multimodel representation learning.
The authors introduce a large, diverse self-collected video dataset for large-scale pretraining and plan to make it public. It has a significant impact since large-scale data is crucial for prevalent transformer-based structures and may benefit future works.
The qualitative analysis in Section 5 does offer me a new insight of audio in video.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g. why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice.
I do have some concerns to confirm whether contrastive span training is more advanced for cross-modality representation learning:
In this paper, Table 1 and Table 3 only approves each modality's contributes and audio matters (compared with VL methods). It lacks to compared with other multi-modal methods. In this way, it may be hard to claim that REVERSE is more advanced when considering additional modalities like audio.

Although it may be heavy computation, do authors consider conducting comparison tests on the pretraining dataset to evaluate the influence of each part in the framework qualitatively?

5. Paper rating (pre-rebuttal).

Weak Accept

7. Justification of rating. What are the most important factors in your rating?
Please refer to the strengths and weaknesses.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.
Yes

9. Limitations and Societal Impact. Have the authors adequately addressed the limitations and potential negative societal impact of their work? Discuss any serious ethical/privacy/transparency/fairness concerns here. Also discuss if there are important limitations that are not apparent from the paper.
yes, this paper provides detailed description of the potential social impact and limitation in both paper and the supplymentary.

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?
Dataset contribution claim in the paper. Indicated in the submission form

14. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal).
Weak Accept

15. Final justification (post-rebuttal).
I think the paper is a good work and well-written. Their experiments on the Situated Reasoning and the Action Anticipation are interesting and impressive. Those results show the potential effectiveness of the contrastive span training on advanced and challenging tasks. The rebuttal answered most of my questions although providing more comparisons are better.

Reviewer #2

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
In order to learn multimodal neural script knowledge representations of videos, the authors propose RESERVE, an innovative framework with a new training objective that learns self-supervised representations through three modalities (audio, text, and vision). To prove the effectiveness of the model, the authors conduct several experiments on two downstream tasks. However, this paper is still insufficient and needs to be improved.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.
1. The authors propose RESERVE, a model that learns multimodal script knowledge representations by fusing vision, audio, and text.
2. The authors propose a new contrastive span matching objective, which facilitates self-supervised multimodal representation learning.
3. The authors build YT-Temporal-1B, which is a large, diverse training dataset containing 20 million YouTube videos and 1 billion frames.
4. Experimental results demonstrate the effectiveness of the proposed method.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g. why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice.

From my perspective, this paper is more like leveraging the self-supervised video pre-training to improve the downstream video understanding tasks. Thus the closest work to this paper may be VATT rather than MERLOT, which is also a unified video pre-training framework that contains vision, audio, and text. But the authors do not mention it in the Related Work.
For the proposed span matching objective, I have two questions as follows. 1) In Equation 1, the summation, I understand it takes all the phrase representation w_t into summation, but it is clear that not all w_t has its corresponding masked prediction wˆ_t, as the authors claim that only 25% of them are masked. 2) I wonder that is it proper to employ Cross-Entropy loss here to measure the similarity between the original modality and the masked one, as it is a strong supervised signal compared with the widely used contrastive loss, like MIL-NCE, which may be more appropriate for the self-supervised learning settings.
Since I think the proposed method is actually a video pre-training framework, I have three questions as follows.
I wonder why the authors build a new video pre-training dataset rather than employ the commonly used one, like HowTo100M used by VATT, UniVL and VLM.
I think more downstream tasks should be included. As the authors claim their method performs well in learning representations jointly through three modalities, it is not convincing to finetune it only on two tasks. Other prevalent video understanding missions, like video retrieval and action recognition, should be also included.
I think more baselines should be taken into comparison, especially in TVQA, as there only contains 3 baselines.

5. Paper rating (pre-rebuttal).
Borderline

7. Justification of rating. What are the most important factors in your rating?
The authors have built a large, diverse training dataset containing 20 million YouTube videos and 1 billion frames. However, why the commonly used datasets were not employed was not well-explained. Meanwhile, more downstream tasks should be included to demonstrate the effectiveness of this work.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.
No

9. Limitations and Societal Impact. Have the authors adequately addressed the limitations and potential negative societal impact of their work? Discuss any serious ethical/privacy/transparency/fairness concerns here. Also discuss if there are important limitations that are not apparent from the paper.
N.A.

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?
Dataset contribution claim in the paper. Indicated in the submission form

14. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal).
Borderline Accept

15. Final justification (post-rebuttal).
The authors have addressed my concerns to a certain degree but I am not fully convinced by their arguments. Consequently, I stay with my original review.

Reviewer #3

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
The paper proposes a novel model named RESERVE, learning to represent videos jointly over time and multiple modalities, i.e., audio, subtitles, and video frames. The model is trained over a novel contrastive masked span

learning objective to capture script knowledge across modalities. The authors conduct a number of experiments and the results demonstrate the superior performance of the proposed model.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.

1) Overall, the paper is well organized.
2) The model is trained on a big dataset of 20 million videos.
3) The authors conduct extensive experiments on various settings.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g. why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice.

1) The proposed model seems complex and may be too ``huge'', So showing the running time may be better..

5. Paper rating (pre-rebuttal).

Weak Accept

7. Justification of rating. What are the most important factors in your rating?

Extensive experiments on various settings.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.

No

9. Limitations and Societal Impact. Have the authors adequately addressed the limitations and potential negative societal impact of their work? Discuss any serious ethical/privacy/transparency/fairness concerns here. Also discuss if there are important limitations that are not apparent from the paper.

The authors have discussed the limitations.

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?

Dataset contribution claim in the paper. Not indicated in the submission form

11. Additional comments to author(s). Include any comments that may be useful for revision but should not be considered in the paper decision.

N/A

14. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal).

Weak Accept

15. Final justification (post-rebuttal).

I have read the rebuttals, and my final recommendation is "'weak accept".