Reviewer Comments:

Reviewer: 1

Recommendation: AQ - Publish In Minor, Required Changes

Comments:
The paper presents an audio-visual neural architecture for sound event detection. The proposed scheme is very insteresting and the way audio and visual streams are processed is very novel. The paper is clear and well written. Resutls are reported on 2 datasets and the paper reports also an interesting ablation study to give insights about the behaviour of the model.

Nevertheless, I have some major concerns at high level, related to the message the paper conveys.
1. My first comment is about the title. I see the Authors' point, but if one claims to be introducing a unified learning framework, I think that there should be experiment on mulitple tasks. In my opinion, the experimental analysis reported in the paper is rich and solid. But it cannot be generalized an a unified learning approch. So, my suggestion, is to revise the title and the claims in the introduction of the paper, being more task specific.

2. The experimental results show that the proposed UAVM approach performs rather similarly to the unimodal pipeline. There seems to be a marginal improvement but in most cases it is within the variance of the results. In addition, for AudioSet it performs worse than the feature concatenation. On the other hand, the proposed architecture improves noticeably wrt the state-of-the-art. The fact that with a very simple fusion strategy the proposed modal independent model outperfoms the state of the art is a remarkable result in my opinion. But the proposed modality combination is not bringing any improvement. So, as before, I think that the results are very interesting but the message that the paper seems to convey is, in my opinion, misleading.
One advantage of the proposed model is that the unified network is smaller. But I am not sure it is something relevant. Authors briefly mention it in the result discussion. If it is relevant, maybe Authors should quantify a bit more.

3. The analysis on the joint audio-visual representaiton is nice and rather convincing. The point is that the single-modal fusion provides the same results. So, very nice, instructive and well done but not very useful at the end.

I do not know if this changes are considered minor or major or not possible. In my opinion it is a matter of revisiting a bit the way the message is conveyed, so I would say they are minor.
At least this is my view.

Thank you very much for reporting results on multiple runs.

Additional Questions:
1. Is the topic appropriate for publication in this transaction?: Yes

2. Is the topic important to colleagues working in the field?: Yes

Explain:

3. How would you rate the technical novelty of the paper?: Very Novel

Explain: The way audio and video features are combined is very novel.

4. How would you rate the English usage?: Satisfactory

6. Rate the references: Satisfactory

Reviewer: 2

Recommendation: AQ - Publish In Minor, Required Changes

Comments:
Clarity

The paper was easy to read and follow. The model is explained nicely -- simply yet giving all the necessary details. The experimentation section gives all the necessary information, then summarizes the findings concisely and informatively. The ablation analysis was also relevant and interesting to read.

Quality

The paper follows the best standards of designing the model and conducting experiments.

Relation to prior publications

While the citations in general are fine, https://openreview.net/forum?id=fXorxxbDvO seems related: it explores the robustness to the missing modality.

Other

- I was confused about the comment on lines 43-44 that MBT uses pretrained Imagenet features. Isn't it the case for UAVM too as mentioned on page 2 line 52 right?
- Some of the results show the overlapping confidence intervals. You cannot draw conclusions if either of them is performing better in this case.

Additional Questions:
1. Is the topic appropriate for publication in this transaction?: Yes

2. Is the topic important to colleagues working in the field?: Yes

Explain:

3. How would you rate the technical novelty of the paper?: Novel Enough for Publication

Explain: The paper introduces a transformer-based model that is able to take two modalities (audio and video) and encode them into the same output space. This is a novel approach to the task at hand and it deserves exploration.

4. How would you rate the English usage?: Satisfactory

6. Rate the references: Satisfactory

Reviewer: 3

Recommendation: R - Reject (Paper Is Not Of Sufficient Quality Or Novelty To Be Published In This Transactions)

Comments:
This paper proposed a unified audio-visual model (UAVM) for audio/video event classification. The work addresses fundamental problems in machine learning. However, several issues are still unclear and therefore require significant revision.

Novelty

+ It claims that "existing works only unify a part of the model components while the proposed system unifies almost everything." But from Fig.1, the proposed model also seems to need a modal-specific feature extractor and modal-specific Transformer, and only the upper layer shares the weights. In other words, it also looks like only unifying a part of the model; therefore, the novelty is unclear.

+ If the model-specific Transformer is only optional, then the experimental results shall justify the performance with and without the modal-specific Transformer.

+ Furthermore, it still uses separate audio-based and video-based predictions and then a fusion layer at last. So except for the part that both modalities use the same architecture, both modality seems still work independently. So the statement "the difference between a unified model and a modal-independent model remains unclear" is unfortunately still valid.

Related Works

There are other existing works that proposed a unified model, such as the one by Ao et al., "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing." Although the modality is different, I think they may discuss the similarity and originality in terms of the Transfomer-based unified model.

Experimental Results

+ Despite several related works on the unified model introduced in the Introduction (such as [14]-[17]), unfortunately, none of those models are included in the experiments. So whether the proposed method could provide advantages in comparison to existing works is unclear.

+ Compared to the independent model, the performance of the unified model doesn't seem to improve significantly.

Analysis of Representation Space

+ Figure 3D shows that the shared Transformer maps two different inputs to a unified space. But it is still unclear whether both modalities share the same space but are still independent or both modalities also have a similar distribution. For example, will the map show the data in a close location if audio and video have the same content?

Additional Questions:
1. Is the topic appropriate for publication in this transaction?: Yes

2. Is the topic important to colleagues working in the field?: Yes

Explain:

3. How would you rate the technical novelty of the paper?: Novel Enough for Publication

Explain: Several parts are unclear, so I can't justify correctly the novelty.

4. How would you rate the English usage?: Satisfactory

6. Rate the references: Satisfactory