NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Trends in Artificial Intelligence and Machine Learning (TAI911S)

Assessment 4: Paper Review

Sara Vatileni

213018691

Table of content

1. **Summary of the Paper……3-4**
2. **Related Work……4**
3. **Limitations of the Paper……5**
4. **Reference……6**

Paper Reviewed:

Zhang, Hanlei, Hua Xu, Fei Long, Xin Wang, and Kai Gao. 2024. “Unsupervised Multimodal Clustering for Semantics Discovery in Multimodal Utterances.”
In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 18–35. https://doi.org/10.18653/v1/2024.acl-long.2

1. Summary of the Paper

a) Problem Being Addressed

The paper addresses the challenge of discovering semantic structures in **multimodal utterances**, such as those that include both speech and visual information without relying on labelled data. Most existing models rely heavily on supervised learning with annotated datasets, which can be expensive and time consuming to curate. The authors aim to develop an unsupervised method to cluster multimodal utterances based on their semantics, enabling applications like intent recognition, dialogue understanding, and human computer interaction without labelled training data.

b) Main Contribution of the Work

The paper proposes a novel **unsupervised multimodal clustering framework** named UMC (Unsupervised Multimodal Clustering). The core contributions include:

- A joint multimodal representation learning module that combines text, audio, and visual features using contrastive learning to align semantic representations across modalities.
- A semantic clustering mechanism that uses pseudo labels generated from cross modal similarity to iteratively refine cluster assignments.
- A demonstration that the approach achieves state of the art results on multiple benchmark datasets without any supervision.

c) Experimental/Theoretical Results

The authors evaluate their model on several public datasets involving multimodal utterances, such as CMU-MOSI, CMU-MOSEI, and MELD.

Key results include:

- UMC outperforms existing unsupervised and some weakly supervised methods on clustering metrics such as Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and F1-score.
- Ablation studies demonstrate the importance of joint multimodal contrastive learning and the iterative clustering process.
- The model generalizes well across datasets, indicating robustness.

2. Related Work

In the domain of **multimodal representation learning**, several foundational works have shaped the current research landscape. Tsai et al. (2019) proposed a Multimodal Transformer for emotion recognition, which demonstrated the effectiveness of cross modal attention mechanisms. Similarly, Hazarika et al. (2020) introduced the Memory Fusion Network, a powerful model for multimodal sentiment analysis that integrates long term context across modalities. Rahman et al. (2021) extended this work by applying Transformer based multimodal fusion to enhance sentiment prediction across textual, visual, and acoustic streams.

In the field of **contrastive learning**, Chen et al. (2020) developed SimCLR, a simple yet highly effective method for visual representation learning through contrastive loss. He et al. (2020) proposed MoCo (Momentum Contrast), which introduced a momentum encoder for building a large and consistent dictionary for contrastive learning. Building on these principles, Miech et al. (2020) presented MIL-NCE, a framework for learning joint video text embeddings using noise contrastive estimation.

Regarding **unsupervised clustering**, Caron et al. (2018) pioneered Deep Cluster, which iteratively clusters image features and uses them to train a neural network. Asano et al. (2020) improved upon this with a self-labelling approach using prototypical contrastive learning to refine cluster quality. Van Gansbeke et al. (2020) introduced SCAN, a semantic clustering technique that leverages nearest neighbour relationships to assign cluster labels in an unsupervised manner.

Finally, in the area of **multimodal utterance understanding**, Zadeh et al. (2017) proposed the Tensor Fusion Network, a model that explicitly captures intra- and inter-modal dynamics for sentiment and emotion analysis. Liang et al. (2018) followed with a recurrent multistage fusion approach that allows for dynamic integration of modality-specific information, enhancing the understanding of complex multimodal language inputs.

3. Limitations of the Paper

Despite its strong performance, the paper has a few limitations:

- Scalability and efficiency: The model uses iterative pseudo-label refinement and joint contrastive learning, which may not scale efficiently to very large datasets without substantial computational resources.
- Dependence on pre-trained encoders: The method relies on pre trained encoders for example BERT for text, ResNet for vision), which could introduce biases or limitations inherited from those models.
- Evaluation scope: While the results are strong on selected benchmarks, additional experiments on more diverse domains like non-English datasets or real-world applications would improve the generalizability claims.
- Interpretability: Like many deep clustering methods, it may be difficult to interpret why certain utterances are grouped together, limiting transparency in high stakes applications.

REFERENCES

1. Tsai, Y. H. H., Bai, S., Yamada, M., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal Transformer for unaligned multimodal language sequences. *Proceedings of ACL 2019*, 6558–6569. https://doi.org/10.18653/v1/P19-1656

2. Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. *Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131. https://doi.org/10.1145/3394171.3413672

3. Rahman, T., Nasrin, M. S., & Mia, M. A. (2021). Multimodal sentiment analysis using transformer-based fusion. *Procedia Computer Science*, 184, 526–533. https://doi.org/10.1016/j.procs.2021.03.065

4. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 119, 1597–1607.

5. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738. https://doi.org/10.1109/CVPR42600.2020.00975

6. Miech, A., Alayrac, J. B., Smaira, L., Laptev, I., Sivic, J., & Zisserman, A. (2020). End-to-end learning of visual representations from uncurated instructional videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9879–9889.

7. Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. *Proceedings of ECCV 2018*, 132–149. https://doi.org/10.1007/978-3-030-01252-6_8

8. Asano, Y. M., Rupprecht, C., & Vedaldi, A. (2020). Self-labelling via simultaneous clustering and representation learning. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Hyx-jyBFPr

9. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Gool, L. V., & Gool, L. V. (2020). SCAN: Learning to classify images without labels. *European Conference on Computer Vision (ECCV)*, 268–285.

10. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *Proceedings of EMNLP 2017*, 1103–1114. https://doi.org/10.18653/v1/D17-1115

11. Liang, P. P., Zadeh, A., Morency, L. P., & Salakhutdinov, R. (2018). Multimodal language analysis with recurrent multistage fusion. *Proceedings of EMNLP 2018*, 150–161. https://doi.org/10.18653/v1/D18-1014