CHALLENGES OF MULTI-MODAL CORESET SELECTION FOR DEPTH PREDICTION

Viktor Moskvoretskii
Skoltech, HSE University
vvmoskvoretskii@gmail.com

Narek Alvandian
Independent Researcher
naalvandyan@gmail.com

ABSTRACT

Coreset selection methods are effective in accelerating training and reducing memory requirements but remain largely unexplored in applied multimodal settings. We adapt a state-of-the-art (SoTA) coreset selection technique for multimodal data, focusing on the depth prediction task. Our experiments with embedding aggregation and dimensionality reduction approaches reveal the challenges of extending unimodal algorithms to multimodal scenarios, highlighting the need for specialized methods to better capture inter-modal relationships.

1 INTRODUCTION

Data for modern deep learning systems require gigabytes and even terabytes of storage (Russakovsky et al., 2015) (Schuhmann et al., 2022) and substantial computational resources for training. One technique to address these computational challenges is coreset selection (Iyer et al., 2021b; Coleman et al., 2020; Chen et al., 2012), which aims to identify minimal subsets of training data that maintain model performance. However, many real-world applications, such as medical diagnosis (Salvi et al., 2024) or autonomous vehicles (Cui et al., 2022; Yeong et al., 2021; Caesar et al., 2020) require processing multiple modalities of data simultaneously. These multimodal scenarios amplify computational demand and introduce new challenges, as traditional coreset selection methods cannot be applied directly. In this work, we extend one of the SoTA coreset selection techniques to handle multimodal data (Zhou et al., 2023). Through extensive experimentation on depth prediction tasks, we demonstrate the limitations of current approaches and the need for specialized multimodal coreset selection methods for better modeling inter-modal relationships. We provide code for reproducing experiments.[1]

2 METHOD

We adapt the Data Quantization (Zhou et al., 2023) method, to the multimodal setting. The goal is to select a representative subset S that retains the diversity and informativeness of the original dataset. Let $D = \{(\{x_i^m\}_{m=1}^M, y_i)\}_{i=1}^N$ denote a dataset with N samples and M modalities, where x_i^m represents the features of the i-th sample for the m-th modality, and y_i is the corresponding target. For ease of notation, x_i will further denote a multimodal tuple $\{x_i^m\}_{m=1}^M$.

Following previous approaches, we employ a submodular gain function (Iyer et al., 2021a) and generalize it to multimodal data. Denoted $P(x_i)$, to measure the importance of a multimodal sample x_i in maximizing the retained information. The gain of adding x_i to the current subset S_{i-1} is:

$$P(x_i) = \sum_{p \in S_{i-1}} ||f(p) - f(x_i)||^2 - \sum_{p \in D \setminus S_{i-1}} ||f(p) - f(x_i)||^2,$$

where $f(x)$ is the embedding of a **multimodal** sample x, S_{i-1} is the current subset of size $i-1$, and $D \setminus S_{i-1}$ represents the remaining samples.

The data set D is divided into non-overlapping bins $\{S_1, S_2, \ldots, S_N\}$ through recursive selection by maximizing submodular gain $x_k \leftarrow \text{argmax } P(x)$ with $x \in D \setminus \bigcup_{j=1}^{n-1} S_j$.

[1] https://github.com/VityaVitalich/MultiModalCoreset

Method	Aggregation	Dimension	Val RMSE, % ↑	Val Loss, % ↑
Full Dataset	-	-	100.00	100.00
Random Coreset	-	-	50.23	46.33
Coreset	Concat	301824	49.08	47.41
	Mean	768	**51.54**	47.90
	Sum	768	51.11	52.12
Coreset w/ PCA	Concat	512	47.73	45.43
	Concat	1024	55.93	50.00
	Concat	2048	47.90	43.23
	Concat	4096	44.00	36.50
Coreset w/ UMAP	Concat	512	49.29	49.23
	Concat	1024	50.53	45.27

Table 1: Percentage of quality retained relative to the Full Dataset for Validation RMSE and Validation Loss, evaluated after training with coresets selected using each method.

3 Experimental Setting

Dataset: We use the CLEVR dataset (Johnson et al., 2016), where multimodal input consists of RGB image and semantic mask, the target is a depth map obtained from Omnidata (Eftekhar et al., 2021).

Models: We employ MultiMAE both to extract a multi-modal embedding $f(x_i)$, and to perform depth estimation from RGB + semantic mask. Embeddings $f(x_i)$ of a multimodal sample x_i are obtained from MultiMAE's transformer encoder. The input and output adapters were trained following (Bachmann et al., 2022) and (Ranftl et al., 2021) respectively. We use Root Mean Square Error (RMSE) as the target metric for depth estimation. Refer to Appendix A for more details.

Coreset Selection: All coresets are 20% from the original dataset, obtained with $N = 20$. We evaluated the following baselines: **Full**: Complete dataset used for training as a reference. **Random Coreset**: A random 20% subset. **Token Aggregation**: Concatenation, mean or sum of embeddings. **Dimensionality Reduction**: We also apply PCA and UMAP (McInnes et al., 2020) to concatenation of tokens before coreset selection.

4 Results & Discussion

Our results in Table 3 show that coreset selection methods lead to a 50% performance drop compared to the full data set and an improvement so minor over the random coreset that it could be due to chance, see Appendix A for raw RMSE values. The best performance is achieved with PCA (1024 features), but the improvement is incremental and UMAP shows no consistent gain.

We hypothesize that these results are due to the large dimensionality of the embeddings, and euclidean distance not being informative enough to separate similar and dissimilar objects. We employed dimensionality reduction to extract the most "meaningful" parts of the embedding, which did not help in our setting. A direction worth exploring would be to try another model for extracting multi-modal embeddings.

5 Conclusion

We address the challenge of selecting multimodal coresets, essential for modern applications, and present an adaptation of the SoTA coreset selection method to the multimodal setting. Our results highlight the need for further exploration of multimodal coreset selection techniques.

REFERENCES

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders, 2022. URL https://arxiv.org/abs/2204.01678.

Holger Caesar, Varun Kumar Reddy Bankiti, Alex Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. pp. 11618–11628, 06 2020. doi: 10.1109/CVPR42600.2020.01164.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding, 2012. URL https://arxiv.org/abs/1203.3472.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning, 2020. URL https://arxiv.org/abs/1906.11829.

Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23:722–739, 02 2022. doi: 10.1109/TITS.2020.3023541.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2021.

Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pp. 722–754. PMLR, 2021a.

Rishabh Iyer, Ninad Khargonkar, Jeff Bilmes, and Himanshu Asnani. Submodular combinatorial information measures with applications in machine learning, 2021b. URL https://arxiv.org/abs/2006.15412.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, October 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Massimo Salvi, Hui Wen Loh, Silvia Seoni, Prabal Datta Barua, Salvador García, Filippo Molinari, and U. Rajendra Acharya. Multi-modality approaches for medical support systems: A systematic review of the last decade. *Information Fusion*, 103:102134, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102134. URL https://www.sciencedirect.com/science/article/pii/S1566253523004505.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21:2140, 03 2021. doi: 10.3390/s21062140.

Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. Dataset quantization, 2023. URL https://arxiv.org/abs/2308.10524.

Method	RMSE	Validation Loss	Training Loss
Full Dataset	0.0033	0.0015	0.0030
Random Coreset	0.0065	0.0033	0.0055
Coreset, Full L2	0.0066	0.0032	0.0058
Coreset, Mean Cosine	0.0063	0.0032	0.0057
Coreset, Sum Cosine	0.0064	0.0029	0.0055
PCA (512)	0.0068	0.0033	0.0066
UMAP (1024)	0.0065	0.0033	0.0062
UMAP (512)	0.0066	0.0031	0.0064
PCA (1024)	0.0059	0.0030	0.0055
PCA (2048)	0.0068	0.0035	0.0057
PCA (4096)	0.0074	0.0042	0.0058

Table 2: Comparison of RMSE, Validation Loss, and Training Loss across methods.

A Technical Details

Training was performed using the Adam optimizer with learning rate 5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.99$, for 40 epochs with batch size 128, no weight decay and cosine annealing scheduler. The training was conducted on a single NVIDIA A100 GPU.

B Acknowledgements

The authors thank Andrei Filatov for the initial idea of this project.