Measuring Modality Utilization in Multi-Modal Neural Networks

Saurav Singh
Dept. Elect. & Microelectronic Eng.
Rochester Institute of Technology
Rochester, NY, USA
ss3337@rit.edu

Panos P. Markopoulos
Dept. Elect. & Comput. Eng. and Dept. Comput. Science
The University of Texas at San Antonio
San Antonio, TX, USA
panagiotis.markopoulos@utsa.edu

Eli Saber
Dept. Elect. & Microelectronic Eng.
Rochester Institute of Technology
Rochester, NY, USA
esseee@rit.edu

Jesse D. Lew
Dept. Computer Science
New York University
New York, NY, USA
jl8429@nyu.edu

Jamison Heard
Dept. Elect. & Microelectronic Eng.
Rochester Institute of Technology
Rochester, NY, USA
jrheee@rit.edu

***Abstract*—Multimodal data provides information from different sensor types about the same underlying phenomenon and enhances machine learning performance. However, neural networks trained end-to-end on all the modalities tend to rely mostly on one of the most dominant modalities. The black box nature of neural networks makes it difficult to assess the reliance of the network on various modalities. This work presents a novel modality utilization metric that quantifies the network reliance on different modalities. The proposed metric is validated on NTIRE-21 (classification problem) and MCubeS (image segmentation problem) datasets. The modality utilization metric contributes towards the explainability of multimodal neural networks and offers great utility in the field of multimodal data fusion.**

***Index Terms*—data fusion, feature importance, multimodal, modality utilization**

I. Introduction

Information about the same phenomenon can be acquired by different sensing modalities, which may capture complementary and redundant information. Each modality in such multimodal data can give optimal information under various conditions. Fields such as aerospace for aerial satellite imagery [1]–[5], autonomous vehicles for image segmentation [6] (see Figure 1), human factors research to estimate human states [7] rely on data from multiple sensing modalities to observe the same phenomena. Thus, there is a need to understand how current multimodal deep-learning data fusion methods utilize each modality during inference.

There are three primary fusion paradigm types: signal/data level, feature level, and decision level [8]. Signal level fusion combines the raw input data prior to inclusion in a machine-learning model [9]. Feature-level data fusion occurs within the machine-learning mode/network architecture [10] (e.g., see Figure1). Decision level fusion makes unimodal (single modality) decisions which it then combines at the inference

This material is based upon work supported in part by the Department of Defense under award HM04761912014 and the Air Force Office of Scientific Research (AFOSR) under award FA9550-20-1-0039.

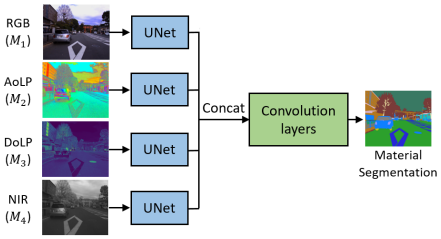

Figure 1. Multimodal Deep Network architecture used for material segmentation with MCubeS dataset [6].

layers [11]. This work mainly focuses on feature level fusion, where end-to-end training is typically employed [12], [13].

Recent research has shown, however, that feature level fusion may place most emphasis on a single (dominant) modality, largely ignoring the rest. In other cases, a mismatch in the optimization hyperparameters for each modality causes end-to-end multimodal trained networks to perform worse than their unimodal counterparts [13], [14]. Thus, knowing to what extent the network utilizes each of the available data modalities provides explainability on the network's operation. This information can help us modify the network to best leverage all modalities or decide to remove lowly utilized modalities, since sensing across multiple modalities comes at some implementation cost/complexity. Works such as Q.T. Truong et al. [15] assessed modality importance by examining the unimodal performances. However, unimodal performances are not necessarily indicative of the network's modality utilization in the multimodal case, as discussed in Section III-C. Therefore, there is a clear gap in the literature on network reliance and modality utilization for multimodal data.

The main research question that we address in this paper is: *How can we quantify the utilization of a modality by the network?* To that end, we introduce a new *modality utilization* (MU) metric. This metric was inspired by the permutation

feature importance metric [16], [17], which quantifies how much a network uses a single feature. Feature utilization metrics are infeasible for fusion of multi-modal images, due to the use of sophisticated convolutional neural networks. Thus, this work expands a permutation-based feature utilization approach to a multi-modal data fusion domain.

The proposed modality utilization metric is experimentally assessed on NTIRE-21 (a classification problem) [3] and MCubeS (an image segmentation problem) [6] datasets. The results show that the proposed metric can quantify the network's reliance on a modality. This contributes to the explainability aspect of a multimodal machine learning model. It also advances the field of data fusion by giving researchers a tool to evaluate a network's reliance on various modalities. The rest of the paper is organized as follows: Section II proposes the modality utilization metric (MU). Section III presents the experimental design, datasets used, experimental results, and discusses the findings, and Section IV concludes this paper.

II. Methodology

Inspired by permutation feature importance [16] [17], the presented method determines modality utilization for a modality MU_i. Given a trained network model F_θ and test dataset $\mathcal{D}_{test}$ with M input modalities, modality utilization MU_i is computed by breaking the association between the input modality M_i and the output label Y and calculating the resulting loss on the testing set.

The association between a modality M_i and the output label Y is broken by permuting or shuffling the corresponding modalities' (M_i) samples randomly, while keeping the remaining modalities' (M_j, $j \neq i$) samples the same, as shown in figure 2. Let independent samples from the testing dataset of the form $\mathcal{D}_{test} = (Y, X_1, X_2, \cdots, X_M)$ be

$$\mathcal{S}^{(a)} = (Y^{(a)}, X_1^{(a)}, X_2^{(a)}, \cdots, X_M^{(a)}),$$
$$\mathcal{S}^{(b)} = (Y^{(b)}, X_1^{(b)}, X_2^{(b)}, \cdots, X_M^{(b)}).$$

A new testing dataset $\mathcal{D}_i$ is then generated, where samples of the i^{th} modality are permuted ($X_i^{(a)}$, $X_i^{(b)}$) as

$$\mathcal{S}^{(b)}_{permuted,i} = (Y^{(b)}, X_1^{(b)}, X_2^{(b)}, \cdots, X_i^{(a)}, \cdots, X_M^{b}). \quad (1)$$

Let the loss of model F_θ be L_{test} during inference with dataset $\mathcal{D}_{test}$ and L_i during inference be with the permuted dataset $\mathcal{D}_i$, where samples of i$^t h$ modality ($X_i^{(a)}$, $X_i^{(b)}$) are permuted:

$$L_{test} = \mathbb{E}\,\mathcal{L}\{F_\theta, (Y^{(b)}, X_1^{(b)}, X_2^{(b)}, \cdots, X_M^{(b)})\}, \quad (2)$$

$$L_i = \mathbb{E}\,\mathcal{L}\{F_\theta, (Y^{(b)}, X_1^{(b)}, X_2^{(b)}, \cdots, X_i^{(a)}, \cdots, X_M^{b})\}. \quad (3)$$

The modality utilization of the i^{th} modality (MU_i) for the model F_θ can be computed by observing the error between the loss during inference with the original dataset $\mathcal{D}_{test}$ and the permuted dataset $\mathcal{D}_i$. Modality utilization of the i^{th} modality (MU_i) is then defined as:

$$MU_i = L_1 - L_{test}. \quad (4)$$

Algorithm 1 formalizes method to compute modality utilization (MU), which takes a given trained network model F_θ and a multimodal test set $\mathcal{D}_{test}$. Line 1 computes the model prediction loss L_{test} (eq. 2). Then, for each modality M_i, permute corresponding test set samples $\mathcal{D}_{test}$ (Line 3) and compute the resulting prediction loss (Line 4). Then, Line 5 computes the modality utilization (MU_i) using eq. 4. Once all of the MUs have been determined, Line 6 computes the normalized percentages of each MU_i.

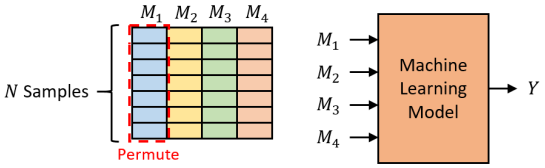

Figure 2. Permuting/shuffling samples of a modality M_i in the dataset to break the association between input modality M_i and the output label Y.

Algorithm 1: Compute Modality Utilization

```
Initialize network model F_θ, and multi-modal test set D_test;
Compute model prediction loss L_test, Eq. 2;
for each modality M_i do
    Randomly permute the samples of modality M_i
      while keeping the modalities M_j, j ≠ i
      unchanged, Eq. 1;
    Compute model prediction loss L_i with permuted
      modality M_i, Eq. 3;
    Compute loss-based Modality Utilization (MU_i)
      using Eq. 4, MU_i = L_i − L_test;
end
Compute normalized percentages for each MU_i.
```

III. Experiments

A. Datasets and Networks

The proposed modality utilization method is validated on two datasets: (i) NTIRE-21 image classification dataset [3], and (ii) MCubeS material segmentation dataset [6].

The NTIRE-21 dataset is an aerial imagery dataset that consists of two modalities, Electro-optical (EO) and Synthetic Aperture Radar (SAR). The dataset presents a classification problem with 10 classes. The EO modality contains more information on a clear day, while the SAR modality is more useful on a cloudy day as SAR can penetrate clouds while EO cannot [3]. This presents an interesting problem as a trained network tends to rely mostly on the EO modality. The network used for the NTIRE-21 dataset consists of two branches (one for each modality) with ResNet18 as the backbone of the feature extractor. The extracted features are fused by concatenating the features from each modality (feature level fusion) and fed to the decision layer which consists of fully connected layers for class prediction. The network was trained using Adam optimizer with 0.001 as the learning rate over 250 epochs and the model with the best test performance is

Table I
PERFORMANCE AND MODALITY UTILIZATION (MU) FOR THE TWO DATASETS COMPUTED USING ALGORITHM 1.

Dataset	Experiment	Modality	Performance	Modality Utilization (MU) (%)			
			Accuracy (%)	EO		SAR	
NTIRE-21	Unimodal	EO	97.5	100.0		-	
	Unimodal	SAR	84.9	-		100.0	
	Multimodal	EO-SAR	97.8	99.59		0.40	
			mIoU	RBG	AoLP	DoLP	NIR
MCubeS	Unimodal	RGB	0.318	100.0	-	-	-
	Unimodal	AoLP	0.266	-	100.0	-	-
	Unimodal	DoLP	0.262	-	-	100.0	-
	Unimodal	NIR	0.270	-	-	-	100.0
	Multimodal	RGB-AoLP-DoLP-NIR	0.374	34.5	19.0	30.9	15.6
	Multimodal	AoLP-DoLP-NIR	0.351	-	67.3	21.0	11.7

used to validate the proposed MU metric. Since the dataset is imbalanced, 624 samples (number of samples in the class with the least number of samples) were used from each class where 524 samples from each class were used as the training set and 100 samples from each class were used as the test set.

The MCubeS dataset is a material segmentation dataset with street scenes that consist of four modalities: colored image (RGB), angle of polarization (AoLP), degree of polarization (DoLP), and near-infrared (NIR). The dataset presents an image segmentation problem with 20 possible categories. Since RGB does not provide enough information to predict the material of a surface, other modalities enhance the prediction capabilities of a network. The network used for the MCubeS dataset consists of four branches (one for each modality) with UNet as the backbone of the feature extractor. The extracted features are fused by concatenating the features from each modality (feature level fusion) and fed to the decision layer which consists of 2-d convolutions layers for image segmentation. The network was trained using SGD optimizer with 0.05 as the learning rate and 0.9 as the momentum over 1000 epochs and the model with the best test performance is used to validate the proposed MU metric. MCubeS dataset consists of 500 samples where 302, 96, and 102 samples are in the training set, validation set and test set, respectively.

B. Results

The network's performance is observed first with each modality separately and then using multiple modalities. The classification accuracy and modality utilization for the NTIRE-21 dataset is recorded. The mean intersection over union (mIoU) for image segmentation and the modality utilization for the MCubeS dataset is recorded. The unimodal network performances serve as a baseline for the experiments. The results in Table I show that for the multimodal experiment with the NTIRE-21 dataset, the network learned to rely on the EO modality (the dominant modality) with 99.59% modality utilization. The modality utilization for the MCubeS, however, indicates that the network uses information from multiple modalities, favoring RGB (34.5% MU) and DoLP (30.9% MU) the most. The MCubeS network is also trained without the RGB modality to see the effects of removing the most utilized modality and a drop in performance (mIoU) from 0.374 to 0.351 is observed.

The sensitivity of the proposed modality utilization method is further studied by degrading the quality of the dominant modality (EO) in the NTIRE-21 dataset. This was achieved by blacking out EO images with a probability of 0%, 25%, 50%, 75%, and 100%. The network is trained from scratch each time while modality utilization and classification accuracy are observed. Figure 3 shows that the network mostly utilizes EO with 0% blackout (also shown in Table I). As the EO blackout increases, the network starts to utilize SAR modality more and more until it only relies on SAR at 100% EO blackout. A drop in the classification accuracy is also observed with an increase in EO blackout, which was expected. This validates the utility of the proposed modality utilization metric as it gives more insight into what modality the network is relying on the most under different conditions and contributes towards the explainability of the black box neural networks.

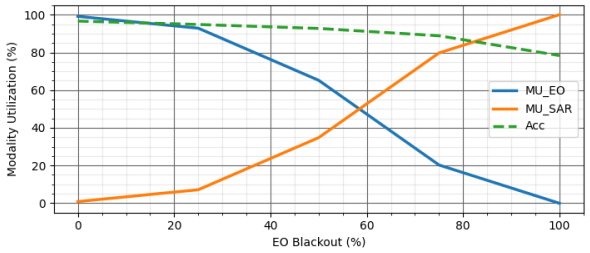

Figure 3. The Modality Utilization Scores and Classification Accuracy by EO Modality Blackout %.

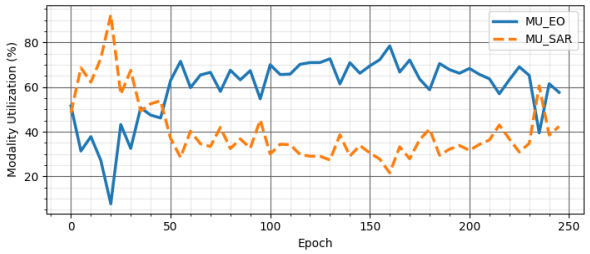

Figure 4. Modality utilization (MU) with 50% EO blackout by training epoch on the NTIRE-21 dataset.

The evolution of the modality utilization metric while training the network on the NTIRE-21 dataset was also observed. The utilization of EO and SAR modalities is observed with nearly 50% EO blackout in Figure 3. Figure 4 shows the evolution of MU with 50% EO blackout while training the network. The proposed modality utilization method helps in tracking the network reliance on the different modalities. It is a useful metric to influence the training process and favor network reliance on one or the other modality.

Redundant information in multiple modalities can affect the network reliance on different modalities. The proposed MU metric allows us to study the effects of redundant information in multiple modalities. A network was trained on NTIRE-21 dataset with two duplicated EO modalities instead of EO and SAR modalities. This ensures the redundant information in the two modalities. The results show that in such a scenario, the modality utilization is heavily influenced by the network's random initialization, shown in Figure 5.

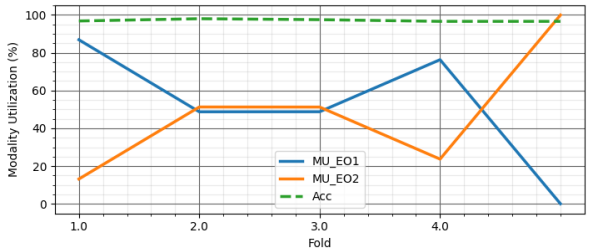

Figure 5. Effects of different network initialization with perfect information redundancy on modality utilization (MU) and classification accuracy.

C. Discussion

Computing modality utilization can help with the explainability of a multimodal fusion network and help in assessing the model reliance on certain modalities. We observe that, as expected, the modality utilization of the dominant modality decreases as the blackout in the dominant modality increases. This was validated by adding blackouts to the EO modality in NTIRE-21 dataset (see Figure 3). As the probability of blackout in EO (dominant modality) is increasing, the MU of EO is decreasing. This suggests that the proposed MU metric is able to give an estimated measure of the true utilization of a modality by a network.

The unimodal performances of the network give an insight into how much information that modality contains compared to other modalities. Higher performance by a modality indicates the modality contains information more appropriate for classification/segmentation. According to Table I, the modality with the highest unimodal performance will also have the highest MU value in the multimodal setting. However, this is not true in the MCubeS dataset. That is, modality performance in unimodal setting does not necessarily reflect the modality importance/utilization in a multimodal setting.

IV. Conclusions

In conclusion, this paper presented a modality utilization metric for multimodal data fusion applications which was inspired by permutation feature importance. The proposed metric was validated using NTIRE-21 (classification problem) and MCubeS (image segmentation problem) datasets. The behavior of the metric was studied by carefully designed experiments where noise/blackouts were added to the dominant modality and expected MU behavior was observed. The future work will consist of using the proposed metric to influence the training of a multimodal network to favor network reliance on one or more modalities. The modality utilization metric offers great utility in fields such as autonomous driving with multimodal data fusion and contributes towards the explainability of the black box multimodal network.

References

[1] M. Sharma, M. Dhanaraj, S. Karnam, D. G. Chachlakis, R. Ptucha, P. P. Markopoulos, and E. Saber, "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE J. of Sel. Topics in App. Earth Observ. and Remote Sens.*, vol. 14, pp. 1497–1508, 2021.

[2] M. Sharma, P. P. Markopoulos, and E. Saber, "YOLOrs-lite: A lightweight CNN for real-time object detection in remote-sensing," in *2021 IEEE Int. Geosci. and Remote Sens. Symp. IGARSS*, Brussels, Belgium, July 2021, pp. 2604–2607.

[3] "NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results," in *Proc. - 2021 IEEE/CVF Conf. on Comput. Vision and Patt. Recogn. Workshops, CVPRW 2021*. IEEE Computer Society, Jun. 2021, pp. 691–700.

[4] E. P. Blasch, U. Majumder, T. Rovito, and A. K. Raz, "Artificial intelligence in use by multimodal fusion," in *2019 22th Int. Conf. on Info. Fusion (FUSION)*, 2019, pp. 1–8.

[5] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, and J. Li, "A survey of multimodal sensor fusion for passive RF and EO information integration," *IEEE Aero. & Elect. Sys. Mag.*, vol. 36, no. 7, pp. 44–61, 2021.

[6] Y. Liang, R. Wakaki, S. Nobuhara, and K. Nishino, "Multimodal material segmentation," in *Proc. of the IEEE/CVF Conf. on Comput. Vision and Patt. Recogn. (CVPR)*, Louisiana, June 2022, pp. 19 800–19 808.

[7] J. Heard, R. Heald, C. E. Harriott, and J. A. Adams, "A diagnostic human workload assessment algorithm for supervisory and collaborative human-robot teams," *ACM Trans. on Human-Robotic Interact.*, vol. 8, no. 2, pp. 1–30, 2019.

[8] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fusion*, vol. 57, pp. 115–129, 2020.

[9] J. Chakraborty and M. Stolinski, "Signal-level fusion approach for embedded ultrasonic sensors in damage detection of real RC structures," *Mathematics*, vol. 10, no. 5, 2022.

[10] S. Liu, Y. Zheng, Q. Du, L. Bruzzone, A. Samat, X. Tong, Y. Jin, and C. Wang, "A shallow-to-deep feature fusion network for VHR remote sensing image classification," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 60, pp. 1–13, 2022.

[11] J. Li, T. Qiu, C. Wen, K. Xie, and F.-Q. Wen, "Robust face recognition using the deep C2D-CNN model based on decision-level fusion," *Sensors*, vol. 18, no. 7, 2018.

[12] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. on Patt. Analysis and Mach. Intell.*, vol. 41, no. 2, p. 423–443, 2019.

[13] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. of App. Earth Observ. and Geoinf.*, vol. 112, no. 102926, 2022.

[14] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *Proc. of the 39th Int. Conf. on Mach. Learn. (ICML)*, vol. 162. PMLR, July 2022, pp. 24 043–24 055.

[15] Q.-T. Truong, A. Salah, T.-B. Tran, J. Guo, and H. W. Lauw, "Exploring cross-modality utilization in recommender systems," *IEEE Intern. Comput.*, vol. 25, no. 4, pp. 50–57, 2021.

[16] L. Breiman, "Random forests," *Machine Learn.*, vol. 45, pp. 5–32, 2001.

[17] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of machine learning research: JMLR*, vol. 20, 2019.