

Supplementary Materials: MB2C: Multimodal Bidirectional Cycle Consistency for Learning Robust Visual Neural Representations

Anonymous Authors

1 THE CONCEPT OF CHANCE LEVEL

The "chance level" in experiments serves as a benchmark for assessing whether the observed results are due to the specific intervention or merely random occurrences. It's essential for setting a baseline expectation when evaluating the effectiveness or impact of an experiment and helps to determine statistical significance beyond chance.

2 IMPACT OF ENCODERS

The performance of the MB2C framework can be significantly impacted by different EEG encoders and image encoders. To this end, several classic methods are selected for comparison. The results are shown in Table 1. EEG encoders included ShallowNet, DeepNet [1], EEGNet [2], and TSConv. TSConv was observed to be the most effective at extracting both temporal and spatial information from EEG, resulting in the best performance. Additionally, ShallowNet and DeepNet were also observed to achieve excellent results in a 200-way classification task, with ShallowNet demonstrating an average top-1 accuracy of 4.55% lower than TSConv, DeepNet demonstrating an average top-1 accuracy of 2.25% lower, and EEGNet demonstrating an average top-1 accuracy of 11.75% lower. We believe that the performance of the MB2C framework could be further enhanced by using other carefully designed EEG feature extractors.

In the case of image encoders, the following models are employed: ResNet-50, which had been pre-trained on ImageNet-1k; ViT-B/16, which had been pre-trained on ImageNet-21k and fine-tuned on ImageNet-1k; and CLIP-ViT-L/14, which had been trained on 400 million image-text pairs. In a 50-way classification task, CLIP achieved an average top-1 accuracy of 50.47%, which is 23.07% higher than ResNet and 19.67% higher than ViT. Even when tested with 200 unseen classes, CLIP demonstrates its superior performance, with an average top-1 accuracy of 28.45% and an average top-5 accuracy of 60.37%, surpassing ResNet by 18.05% and ViT by 10.55%.

3 SUPERCLASSES

In high-level or coarse-grained visual classification tasks, there exist categories that typically encompass multiple images that would be considered different categories at the basic level. These categories share a common label. For example, a superclass could be "animals," which might include various basic categories such as "cat", "lamb", and "goose", even though these basic categories may be widely separated in feature space.

The majority of the 200 novel classes in the test dataset were selectively divided into six superclasses based on conceptual similarity. These superclasses were defined as follows: animals, clothes, food, household, tools, and transportation. The detailed results are listed in Table 2. The partitioning of the classes into superclasses was not only for aesthetic purposes in visualization but also to

investigate whether the model can automatically cluster similar major categories more tightly in the feature space and separate them from different superclasses.

4 EVALUATION METRICS

Inception score (IS) [3]: The Inception Score is a measure used primarily in the field of evaluating the performance of generative models.

The Inception Score is introduced as a way to quantify the quality of the generated images. It is based on the Inception-v3 classifier [4], which is a pre-trained deep neural network developed for image classification.

$$IS = \exp \left(\mathbb{E}_{x \sim p_g} [D_{KL}(p(y|x} || p(y))] \right), \quad (1)$$

D_{KL} represents the Kullback-Leibler divergence, $P(y|x)$ is the predicted distribution of categories given the image x , and $P(y)$ is the marginal probability of the predicted category distributions across all images.

Frechet Inception Distance (FID) [5]: The Frechet Inception Distance, is a metric that assesses the quality of generated images by comparing the discrepancy in the feature space distributions between synthetic and real images, thus evaluating the performance of generative models.

$$FID = \|\mu_1 - \mu_2\|^2 + \text{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}) \quad (2)$$

μ_1 and μ_2 represent the mean vectors of the feature distributions for generated and real images, respectively, while Σ_1 and Σ_2 are their respective covariance matrices. Tr denotes the trace of a matrix, which is the sum of its diagonal elements.

Kernel Inception Distance (KID) [6]: The Kernel Inception Distance, utilizes the Maximum Mean Discrepancy (MMD) to quantify the disparity between two probability distributions.

$$KID = \max_{\|h\| \leq 1} |h^T(\mu_g - \mu_r)| \quad (3)$$

In this equation, μ_g and μ_r represent the mean embeddings of generated and real images, respectively, in the feature space. h is a vector within the unit ball, with $\|h\| \leq 1$ indicating that the norm of h in the Hilbert space does not exceed 1.

Table 1: Classification accuracy (%) of N -way Top- K with different EEG encoder and Image encoder on ThingsEEG dataset

Type	Method	EEG encoder																					
		Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
		top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
200-way	ShallowNet	17.5	52.5	25.5	51.0	28.5	60.5	31.0	61.5	18.5	46.5	28.5	58.5	20.0	50.0	37.5	64.0	6.5	23.5	25.0	59.0	23.9	52.7
	DeepNet	17.0	46.5	23.0	54.0	24.5	57.0	31.0	67.0	24.0	53.5	27.0	58.5	24.5	54.0	36.0	67.0	26.0	54.5	28.5	68.0	26.2	58.0
	EEGNet	12.0	37.0	6.5	25.5	9.5	29.5	29.5	63.5	4.5	13.5	6.0	27.0	21.5	47.0	35.5	63.5	13.5	36.5	28.5	64.0	16.7	40.7
	TSCov(MB2C)	23.67	56.33	22.67	50.50	26.33	60.17	34.83	67.00	21.33	53.00	31.00	62.33	25.00	54.83	39.00	69.33	27.50	59.33	33.17	70.83	28.45	60.37
50-way	ShallowNet	40.0	82.0	38.0	78.0	46.0	82.0	50.0	88.0	46.0	72.0	64.0	90.0	34.0	76.0	60.0	92.0	20.0	48.0	36.0	88.0	43.4	79.6
	DeepNet	40.0	80.0	46.0	84.0	38.0	82.0	46.0	86.0	48.0	74.0	56.0	84.0	44.0	80.0	50.0	82.0	40.0	80.0	56.0	86.0	46.4	81.8
	EEGNet	20.0	76.0	18.0	44.0	26.0	62.0	42.0	90.0	18.0	46.0	26.0	60.0	30.0	82.0	58.0	88.0	36.0	74.0	44.0	88.0	31.8	71.0
	TSCov(MB2C)	41.33	83.33	38.67	82.67	48.67	84.67	56.00	84.67	39.33	70.00	54.67	86.67	45.33	80.67	68.67	89.33	53.33	89.33	58.67	90.67	50.47	84.20
Type	Method	Image encoder																					
		Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
		top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
200-way	ResNet	6.5	25.0	6.0	27.0	10.5	31.0	17.5	43.0	5.0	16.5	9.5	33.5	13.0	36.5	14.5	48.5	8.5	33.0	12.5	42.0	10.4	33.6
	ViT	16.0	33.5	13.5	35.5	16.0	37.5	17.5	41.0	10.0	35.0	17.5	42.5	20.5	44.5	23.5	53.5	15.5	37.0	28.5	52.0	17.9	41.2
	CLIP(MB2C)	23.67	56.33	22.67	50.50	26.33	60.17	34.83	67.00	21.33	53.00	31.00	62.33	25.00	54.83	39.00	69.33	27.50	59.33	33.17	70.83	28.45	60.37
50-way	ResNet	22.0	54.0	20.0	56.0	32.0	66.0	34.0	74.0	12.0	46.0	26.0	74.0	26.0	58.0	42.0	86.0	20.0	64.0	40.0	68.0	27.4	64.6
	ViT	24.0	64.0	22.0	60.0	22.0	58.0	32.0	74.0	28.0	56.0	30.0	68.0	36.0	62.0	46.0	78.0	32.0	56.0	36.0	82.0	30.8	65.8
	CLIP(MB2C)	41.33	83.33	38.67	82.67	48.67	84.67	56.00	84.67	39.33	70.00	54.67	86.67	45.33	80.67	68.67	89.33	53.33	89.33	58.67	90.67	50.47	84.20

Table 2: To introduce more detailed information of the ThingsEEG test set for zero-shot classification task and its structures inside, we list all the superclasses corresponding to each subclass.

Superclass	classes	label	Superclass	classes	label	Superclass	classes	label	Superclass	classes	label	Superclass	classes	label
animals	antelope	0	clothes	bonnet	1	food	coconut	2	household	television	3	transportation	wheelchair	5
animals	beaver	0	clothes	chaps	1	food	coffee-bean	2	household	treadmill	3	transportation	unicycle	5
animals	cheetah	0	clothes	cleat	1	food	cookie	2	tools	blowtorch	4	transportation	cruise-ship	5
animals	crab	0	clothes	tube-top	1	food	cordon-bleu	2	tools	bottle-opener	4	transportation	ferry	5
animals	eel	0	clothes	coat	1	food	creme-brulee	2	tools	bullet	4	transportation	golf-cart	5
animals	elephant	0	clothes	coverall	1	food	crepe	2	tools	candlestick	4	transportation	gondola	5
animals	flamingo	0	clothes	duffel-bag	1	food	croissant	2	tools	chain	4	transportation	jeep	5
animals	gopher	0	clothes	glove	1	food	cupcake	2	tools	wok	4	transportation	minivan	5
animals	gorilla	0	clothes	headscarf	1	food	dessert	2	tools	vise	4	transportation	sailboat	5
animals	grasshopper	0	clothes	hoodie	1	food	fruit	2	tools	chopsticks	4	transportation	scooter	5
animals	lamb	0	clothes	kneepad	1	food	garlic	2	tools	cleaver	4	transportation	station-wagon	5
animals	piglet	0	clothes	muff	1	food	hamburger	2	tools	dagger	4	transportation	submarine	5
animals	possum	0	clothes	pajamas	1	food	orange	2	tools	fork	4			
animals	rhinoceros	0	clothes	pocket	1	food	onion	2	tools	hammer	4			
animals	turkey	0	clothes	purse	1	food	pear	2	tools	handbrake	4			
animals	bug	0	clothes	sandal	1	household	bench	3	tools	metal-detector	4			
animals	cat	0	clothes	suit	1	household	breadbox	3	tools	music-box	4			
animals	caterpillar	0	clothes	t-shirt	1	household	cd-player	3	tools	pickax	4			
animals	cobra	0	food	top-hat	1	household	chest2	3	tools	pocketknife	4			
animals	crow	0	food	banana	2	household	coffeemaker	3	tools	punch2	4			
animals	dalmatian	0	food	birthday-cake	2	household	crib	3	tools	spatula	4			
animals	dragonfly	0	food	bok-choy	2	household	freezer	3	tools	spoon	4			
animals	pug	0	food	bread	2	household	highchair	3	tools	tongs	4			
animals	eagle	0	food	bun	2	household	lampshade	3	transportation	aircraft-carrier	4			
animals	goose	0	food	calamari	2	household	laundry-basket	3	transportation	bike	5			
animals	panther	0	food	cashew	2	household	nightstand	3	transportation	buggy	5			
animals	pigeon	0	food	cheese	2	household	table	3	transportation	cart	5			

REFERENCES

[1] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggersperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[2] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[3] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[6] Mikołaj Binkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.