Using Human Perception to Regularize Transfer Learning

Supplementary Material

Here, we discuss additional details to the main paper.

1. Datasets

1.1. Psych-ImageNet

- The dataset has 293 known classes in total, excluding other open-set classes to use at later studies.
- There are 40 classes with psychophysical labels, producing a ratio of psychophysically-annotated to original classes as in [10].
- There are 33,548 known training samples in total, and 12,428 samples have corresponding reaction times.
- Each data point has a reaction time, class label, and ImageNet-sized (224x224) image associated with it.
- Reaction times collected (each reaction time is the amount of time to choose a stimulus, given 5 other examples). Responses in this data were collected for class recognition against noisy stimuli.

Each trial was an object-matching task, where the human participant of 5 opposed stimuli to it. Each image was from one of the 293 Psych-ImageNet classes chosen for the task. The participant had to select the object they thought belonged to the top sample or rejected it together, should there be no match. A timer collected the participants reaction time for each question. Best viewed in color. An example of crowd-sourced task pairings from [6] is shown in Fig. 1.

Classes were evenly distributed across trials, as were positive *vs*. negative matches. Likewise, the difficulty of experiments was variable to avoid a ceiling effect, a form of scale attenuation in which the maximum performance measured does not reflect the true maximum of the independent variable.

1.2. Psych-Omniglot

The Psych-Omniglot is a variant on the Omniglot dataset [7] with psychophysical labels collected from the research in [2]. The dataset contains images of handwritten characters from hundreds of typesets, many of which a typical crowd-sourced study participant would be unfamiliar with. The data is augmented with counterpart samples for each image with a deep convolutional generative adversarial network (DCGAN) [4] to increase intraclass variance and the sample size per class — all of which are forms of implicit regularization. An example of crowd-sourced task pairings from [2] is shown in Fig. 2.

In this dataset, human behavioral measurements were gathered as reaction times to stimuli in crowd-sourced experiments. Human participants were presented with two opposing stimuli from the original Omniglot dataset (a Two-Alternative Forced Choice task) and decided whether the two stimuli were the same character in the dataset. The reaction time from the participants was recorded automatically. Broadly speaking about the dataset as a whole, these human reaction times were long on hard pairings, and short on easy character pairings The introduction of this easy *vs*. hard pairing would prove useful for supervised learning tasks.

1.3. Psych-IAM

The dataset is a modification of the IAM dataset [9] with human behavioral measurements on lines of text collected from [5] on about 35% of the dataset (2,152 lines).

In the main text, we report both word error rate (WER) and character error rate (CER) for this dataset. The word error rate is a model's error with respect to the individual word on the line in the dataset, while the character error rate corresponds to the model's fidelity with the human annotator's marks on the individual word.

The reaction time of the annotator to accurately record a character and line was recorded in this annotated dataset [5]. For the main text, we used the reaction times in conjunction with images of the text itself to perform transfer learning OCR tasks with PERCEP-TL Fig. 2.

1.4. Dataset Limitations

We recognize that Psych-ImageNet only contains annotations on 40 of the total 293 classes. While previous psychophysics and machine learning studies suggest that this still remains representative of the entire training distribution [5, 10], reaction times on more classes may potentially yield better results.

Likewise, Psych-Omniglot is a dataset in which the annotations were collected *via* Amazon Mechanical Turk. While the practitioners accounted for systematic errors, no crowd-sourcing study is entirely robust to untrustworthy annotators [11].

Lastly, Psych-IAM, along with the other two datasets, also suffer from limits of overall numbers of available annotations due to academic budget constraints.

2. PredNet Fine-Tuning

Similarly to Blanchard et al. [1], we extract the activations of PredNet Fig. 3 after the convLSTM layers. PredNet works with temporal data. First, we pre-trained it on videos from the KITTI [3] self-driving dataset. We set up the annotated Psych-ImageNet in an order of fixed frames and record the activations of PredNet at the fixed time steps. This repre-

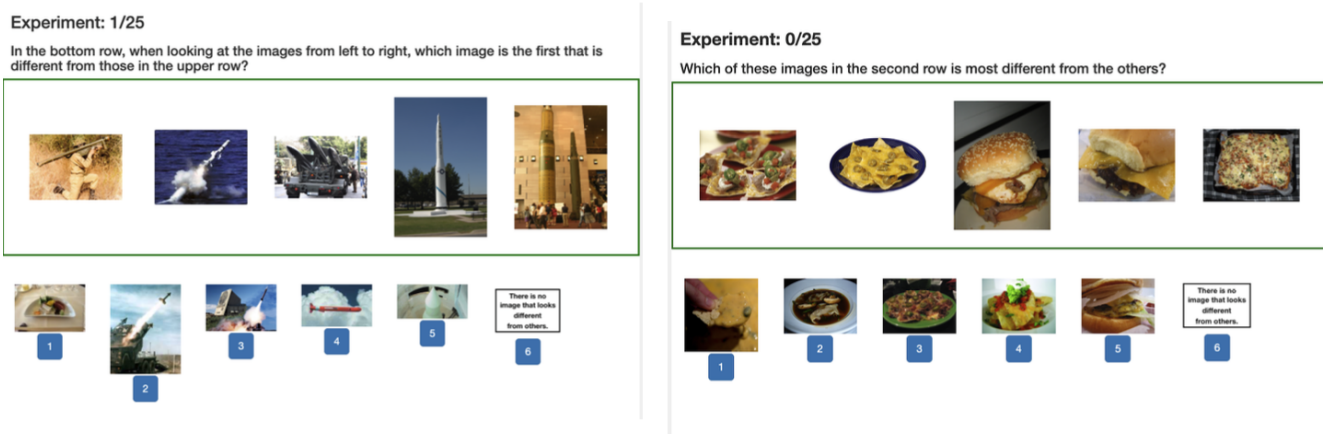

Figure 1. **Crowd-sourced tasks from [6].** The above figure contains two screenshots from the worker data aggregator view in Amazon Mechanical Turk. The image on the left contains an easy example where most annotators answered quickly and accurately; a model that fails to answer in the same way receives a higher penalty. Likewise, the screenshot on the right contains a more difficult class, where a model does not receive as harsh of a penalty for answering incorrectly.

Figure 2. **Crowd-sourced tasks from [2].** An example two-alternative forced choice OCR task as seen from the participant's view. Labels (d) and (f) represent character pairs where the class labels differ; the rest represent the same class pairing. The blurred and noisy images lead to more informative psychophysical labels for operationalization within the machine learning task during training.

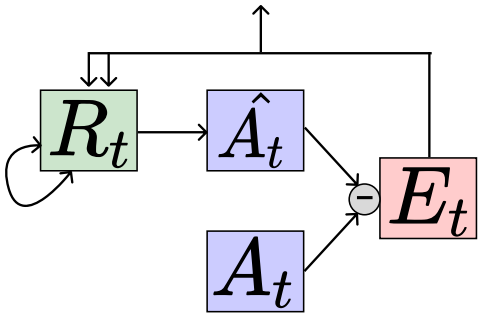

Figure 3. The PredNet[8] architecture.

sentation at each neuron works like a supervised model's neuron in that we can add psychophysical transfer learning to it to regularize the learning representation.

The loss defined by PredNet is as follows:

$$\mathcal{L}_{train} = \sum_t \lambda_t \sum_l \frac{\lambda_l}{n_l} \sum_{n_l} E_l^t \tag{1}$$

where λ_t is a regularizer at the time step, $\frac{\lambda_l}{n_l}$ is a regularizing factor at a given layer in the network, and E_l^t is the error at a time step 2.

Indeed, the loss formulation for PredNet is inherently more complex than cross-entropy loss variations. For brevity, we conducted experiments to understand in which term should the psychophysical regularization variables be used Again, each of the three terms uses a form of ℓ_1-normalization to adjust model learning generalization.

We observe that multiplying psychophysical transfer learning into the layer term λ_t — with the variable $\hat{A}_t$ yields the best results. In other words, after successive outputs of each layered convLSTM, we see performance gains more vividly than any other term within this loss. Furthermore, the regularization effect of psychophysical transfer learning, the softening of sharp gradient turns, pronounces the most at longer time steps on average (*e.g. at steps* > 5). As table Tab. 1 suggests, the loss mostly benefited from psychophysical transfer learning regularization on the outputs between step outputs $\hat{A}_t$ at times t Tab. 2.

This result suggests that predictive coding networks in some way manage the latent ideals encoded in the psychophysical transfer learning data.

While this experiment step was not part of the *model-evaluative*, we believed it important to fine-tune psychophysical transfer learning on a non-traditional loss framework

Psych-ImageNet Method	MAE PredNet
Control	0.59 ± 0.03
ℓ_1	0.62 ± 0.02
ℓ_2	0.61 ± 0.03
Dropout	0.61 ± 0.05
Dropout+ℓ_1	0.61 ± 0.02
$RegularPsych$	0.64 ± 0.04
$RegularPsych$+Dropout	$\mathbf{0.65 \pm 0.02}$

Table 1. On models using $RegularPsych$ as an evaluator, we see improved mean squared error reduction. All models were pre-trained on KITTI and evaluated on the house dataset. We computed error bars using standard error across 5 seeds. *Lower is better.*

Psych-ImageNet		Psych-Omniglot
Parameter	Train Error	Train Error
None	0.12 ± 0.03	0.20 ± 0.05
A_t	0.11 ± 0.04	0.19 ± 0.05
$\hat{A}_t$	0.06 ± 0.02	0.17 ± 0.04

Table 2. The table shows the train errors for each parameter selection of *which* PredNet architecture layer to multiply by the $RegularPsych$ variable. The None column assumes a cross-entropy loss without any modification to the PredNet loss. The input layer A_t shows no significant change in performance, regardless of what the psychophysical annotations are. However, we see a significant reduction in training error when applying $RegularPsych$ to the model prediction logits $\hat{A}_t$.

before conducting experimentation on the relative effects of psychophysical transfer learning on model-evaluative performance.

The pre-training of PredNet and the subsequent transfer to task to a frame-by-frame prediction on the modified Psych-ImageNet allows for the beneficial usage of psychophysical transfer learning. While this case is a niche, it demonstrates the viability of utilizing psychophysical transfer learning in a variety of future neurologically-inspired models.

3. Model Ablations

In Tab. 3, we report ablation results on the PERCEP-TL transfer learning tasks. These show some additional transfer learning movements among different tasks in the experiments. For example, the first row of the figure represents the task shift, where the color of the ψ represents the domain the psychophysical labels were gathered on. Not all domains transfer well, but there exist several domains where transfer learning works naturally.

In this work, it remains apparent that the object recognition task and psychophysical labels from models learned on Psych-ImageNet transfer well to the other domains used in this study. In rows 1 and 3 in Tab. 3, we see the largest gains supported by this. Likewise, the transfer of domains from Psych-IAM character annotation tasks to generic object recognition, in line with commonsense, does not transfer well.

In future studies, we plan to explore different learning paradigms (e.g. reinforcement learning) to expand the results of transfer among domains.

Transfer Task	orig.	+ new	+ %diff ResNet	orig.	+ new	+ %diff VGG	orig.	+ new	+ %diff ViT	orig.	+ new	+ %diff PredNet
$\psi \rightarrow \psi$	0.79	0.81	+1.5%	-			0.83	0.85	**+1.9%**	0.63	0.65	+1.2%
$\psi \rightarrow \psi$	0.74	0.75	+0.4%	0.76	0.76	+0.4%	0.78	0.79	+0.7%	0.65	0.65	+0.1%
$\psi \rightarrow \psi$	0.91	0.92	+0.9%	0.81	0.02	-0.5%	0.86	0.88	+1.2%	0.64	0.65	+1.1%
$\psi \rightarrow \psi$	0.74	0.73	-0.6%	0.76	0.76	+0.2%	0.78	0.77	-0.5%	0.65	0.62	-3.1%
$\psi \rightarrow \psi$	0.91	0.91	+0.4%	0.81	0.81	+0.1%	0.86	0.86	-0.1%	0.65	0.66	+1.2%

Table 3. **Transfer learning % difference table.** With psychophysical transfer learning, performance increases by as much as **1.9**%. Each row represented in this table represents a difference transfer learning task, denoted by ψ corresponding in color with the dataset used. Each trial is the standard error across 5 seeds. *Note: using accuracy as* $1 - CER$ *on Psych-IAM trials in this table. Higher is better.*

References

[1] Nathaniel Blanchard, Jeffery Kinnison, Brandon Richard-Webster, Pouya Bashivan, and Walter J Scheirer. A neurobiological evaluation metric for neural network model search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5404–5413, 2019. 1

[2] Justin Dulay, Sonia Poltoratski, Till S Hartmann, Samuel E Anthony, and Walter J Scheirer. Guiding machine perception with psychophysics. *arXiv preprint arXiv:2207.02241*, 2022. 1, 2

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 1

[5] Samuel Grieggs, Bingyu Shen, Greta Rauch, Pei Li, Jiaqi Ma, David Chiang, Brian Price, and Walter Scheirer. Measuring human perception to improve handwritten document transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[6] Jin Huang, Derek Prijatelj, Justin Dulay, and Walter Scheirer. Measuring human perception to improve open set recognition. *arXiv preprint arXiv:2209.03519*, 2022. 1, 2

[7] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 1

[8] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017. 2

[9] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5 (1):39–46, 2002. 1

[10] Walter J Scheirer, Samuel E Anthony, Ken Nakayama, and David D Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1679–1686, 2014. 1

[11] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, 21(10):736–748, 2017. 1