Understanding Task Representations in Neural Networks via Bayesian Ablation

Andrew Joohun Nam
AI Lab, Princeton University

Declan Campbell
Princeton Neuroscience Institute, Princeton University

Thomas L. Griffiths
Department of Psychology, Princeton University

Jonathan D. Cohen[1]
Princeton Neuroscience Institute, Princeton University

Sarah-Jane Leslie[1]
Department of Philosophy and Center for Statistics and Machine Learning, Princeton University

Abstract

Neural networks are powerful tools for cognitive modeling due to their flexibility and emergent properties. However, interpreting their learned representations remains challenging due to their sub-symbolic semantics. In this work, we introduce a novel probabilistic framework for interpreting latent task representations in neural networks. Inspired by Bayesian inference, our approach defines a distribution over representational units to infer their causal contributions to task performance. Using ideas from information theory, we propose a suite of tools and metrics to illuminate key model properties, including representational distributedness, manifold complexity, and polysemanticity.

Keywords: representation learning; interpretability; neural networks; Bayesian inference

Neural networks have long been used as tools for understanding human cognition (Rumelhart, Hinton, & McClelland, 1986), from minimalist architectures with just 12 learnable weights (Cohen, Dunbar, & McClelland, 1990) applied to cognitive control in the Stroop task (Stroop, 1935), to large-scale language models such as GPT-3 (Achiam et al., 2023) with 175 billion parameters that exhibit human-like cognitive biases and irregularities (Binz & Schulz, 2023; Binz et al., 2024; Lampinen et al., 2024; Webb, Holyoak, & Lu, 2023). As these models grow increasingly complex, however, their underlying representations and processes become more opaque, with mechanistic interpretation restricted to simpler architectures such as linear networks (Saxe, McClelland, & Ganguli, 2019) and attention-only transformers (Olsson et al., 2022). This challenge is particularly pronounced in interpreting latent representations of tasks, especially as language models approach limitless capacity for learning tasks and domains described in natural language (Bubeck et al., 2023; Yu et al., 2023).

In this work, we present a method for exploring task representations using neural ablations to observe the downstream effects on task performance. We define an ablation mask as a binary vector that indicates which representational units to lesion these units by setting their activation values to 0. While traditional ablation studies investigate $P(\text{correct} \mid \text{task}, \text{mask})$, thereby assessing how task performance changes when specific representational units of a model are ablated, our approach instead applies a Bayesian perspective, computing an ablation mask distribution (AMD) to infer which units are most likely to have been used to produce correct responses for a given task, expressed as $P(\text{mask} \mid \text{task}, \text{correct})$. That is, we compute the distribution over possible masks, conditioned on correct task performance. If a specific set of units is crucial for the task, the probability of masking them given success will be low. The ablation mask distribution captures higher-order interactions and complex manifold structures by modeling full statistical dependencies. Thus, our method interprets models without imposing architectural assumptions or constraints.

Beyond the interpretation of individual unit roles, measures that summarize and quantify distributional properties facilitate the quantification of broader structure. For instance, entropy (Shannon, 1948), which measures the concentration of the mask distribution, reveals how localized or distributed a representation is for a given task. This enables exploration of global phenomena within a unified framework for interpreting both micro-level unit functions and macro-level patterns.

We begin by defining the distribution over ablation masks and its relationship to the model's task performance. Next, we demonstrate the utility of our method by applying it to the Integrated Semantics and Control (ISC) model (Giallanza, Campbell, Cohen, & Rogers, 2024), a simple feed-forward multitask neural network trained on human-rated semantic data designed to investigate emergent semantic cognition in context-switching scenarios. We selected the ISC model for its alignment with human responses on measures such as context similarity and for its architectural simplicity, which facilitates the application and validation of novel methods. Using this model, we highlight several information-theoretic analyses enabled by our probabilistic formulation of task representations to quantify network properties such as distributedness of task representations and their manifold complexity, and task representational similarity. We also demonstrate a reverse inference approach that infers the task from the activated representational units as a method to measure polysemanticity. Finally, we discuss the limitations of our method and outline potential future directions to address them.

Supplementary information (SI) are available online[2].

Methods

Integrated Semantics and Control (ISC) model

The Integrated Semantics and Control (ISC) model (Figure 1) is trained on the Leuven Concepts Database (De Deyne & Storms, 2008; Ruts et al., 2004; Storms, 2001), a human-rated

[1] Equal contribution; authors listed alphabetically

[2] https://github.com/andrewnam/isc_ablation_cogsci

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

semantics dataset containing 350 animals and 2,896 features that are grouped into 36 distinct feature classes (Wu & Barsalou, 2009) (see SI Table 1). The model is trained to simultaneously predict the features of a particular animal (item input) and also a subset of its features within a particular feature class (task input). For instance, giving the model the animal "elephant" and the "category" feature class would produce positive outputs only for features relevant to an elephant's category, e.g. "is an animal".

We adopt the architecture described by Giallanza et al. (2024). The inputs—item (animal) and task (feature class)—are represented as one-hot vectors (i.e. a vector of 0s with a single 1), which are mapped to separate embedding spaces: the context-independent representation layer and the task representation layer. The context-independent layer is used to directly predict all features of the animal. It also provides input to the context-dependent layer, along with the task representation, which together form the context-dependent representation. Notably, the task representation—which we apply our ablations to—modulates the context-independent representations by effectively directing the network's attention to the features of the input that are most relevant to the specified task. This context-dependent representation is then used to predict the set of features specified by the task input. We also introduce a null-task during training, represented by a zero-vector embedding and a zero-vector target output, which effectively encourages the model to learn strongly negative output biases and more structured embeddings across the 36 feature classes.

In this paper, we apply our method to the task representation layer of the ISC model. Although using a model with an explicit task representation limits the validity and generality of an approach designed to be applicable even to models without such representations, this choice serves a critical purpose. Starting with a model that has clearly defined task representations allows us to rigorously evaluate a novel approach in a controlled environment, where the model's properties are well understood. By first validating the approach in this setting, we aim to establish a firm foundation for extending these tools to provide new insights in less interpretable systems.

Ablation mask distribution

We define the ablation mask distribution $P(\text{mask} \mid \text{task}, \text{correct})$ as a conditional distribution over binary vectors that mask representational units of a neural network. If the mask value is 0, the representational unit is replaced with zero; if the value is 1, the unit is left unchanged.

In our experiments, we apply these masks to the task representation layer of the ISC model, multiplying the binary mask vector $m \in \{0,1\}^d$ element-wise by the post-activation values of the task representation units $h \in (0,1)^d$, where h follows a sigmoid activation function. We measure model performance conditioned on a task and an ablation mask, $P(\text{correct} \mid \text{task}, \text{mask})$, by applying the mask and taking the feature predictions for all 350 animals. A prediction is mapped to `true` if the predicted feature likelihood in the output layer is ≥ 0.5 and `false` otherwise. Each prediction is compared against the target value in the data to determine whether or not it is correct.

Given the highly skewed distribution of positive and negative feature values in the dataset, we estimate task-mask performance using the geometric mean of the model's sensitivity and specificity:

$$P(\text{correct} \mid \text{task}, \text{mask}) = \underbrace{\sqrt{P(\text{correct} \mid \text{task}, \text{mask}, \text{target} = 1)}}_{\text{sensitivity}} \times \underbrace{\sqrt{P(\text{correct} \mid \text{task}, \text{mask}, \text{target} = 0)}}_{\text{specificity}}$$

Including the null-task described above, which encourages the model to predict 0 for all features, and the geometric mean, which balances evaluation of positive and negative features, results in a 0% "chance" accuracy on all feature classes.

For brevity, we denote the ablation mask as m, the task as t, and the correctness indicator as c. Using this notation, the correctness probability serves as the basis for defining the ablation mask distribution with Bayes' rule:

$$P(m \mid t, c) = \frac{P(t \mid m, c) P(m \mid c)}{\sum_{m'} P(t \mid m', c) P(m' \mid c)}$$

This distribution over ablation masks identifies the subset of causally relevant units that allow the model to successfully perform a specific task, offering a principled framework for interpreting the functional contributions of representational units within neural networks.

While the Bayesian formulation of the correctness metric is mathematically valid, its separation between high and low performance can yield a posterior distribution that is too flat to be useful in practice. For instance, in a task with a baseline accuracy of 50%, a "success" mask achieving 95% accuracy would be sampled only about twice as often as a "failure" mask that performs considerably worse (i.e. reduces performance to chance). This distribution can result in an unbalanced exploration of mask space, potentially making it harder to interpret the functional contributions of different units.

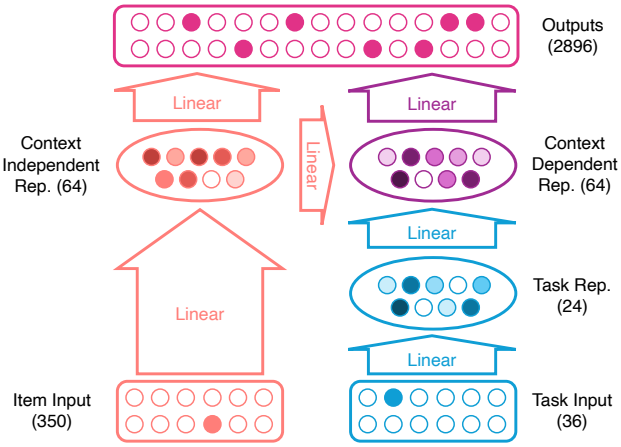

Figure 1: ISC Model. Number of units shown in parentheses. Sigmoid activation function is applied after each linear layer.

Instead, we convert accuracy measures p into odds-ratios $\frac{p}{1-p}$, which amplify performance differences, so that a mask with accuracy $p = 0.95$ is sampled approximately 20 times more often than one with accuracy $p = 0.50$. This aligns naturally with the sigmoid nonlinearity inherent in the model, given by

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad x = \log\left(\frac{p}{1-p}\right).$$

The sigmoid function serves as the inverse of the log-odds transformation, mapping log-odds into probabilities. Focusing on the odds-ratio thus measures the impact of the mask on the input to the sigmoid.

Assuming a uniform prior over the ablation masks, the distribution of the mask given the task and correctness is expressed as:

$$P(m \mid t,c) = \underbrace{\frac{P(c \mid t,m)}{1-P(c \mid t,m)}}_{\text{Likelihood}} \cdot \underbrace{\left(\sum_{m'} \frac{P(c \mid t,m')}{1-P(c \mid t,m')}\right)^{-1}}_{\text{Normalization}}$$

We find that the odds-ratio modification aligns with other measures describing model representation and behavior better than using the standard Bayesian formulation (see SI).

Analyses

One central advantage of the ablation mask distribution is its ability to distinguish between causally relevant and merely incidental activations of representational units. In this section, we apply a suite of information-theoretic measures to quantify key properties of the task representations, including the distributedness of task representations and their manifold complexity, the degree of task polysemanticity, and task representational similarity. All analyses were performed on 10 separately trained instances of the ISC-model.

Entropy

We begin our analyses by considering the entropy $H(m \mid t,c)$ of the ablation mask distribution, which measures the diversity of masks that are sufficient to perform well on a task.

$$H(m \mid t,c) = -\sum_{m} P(m \mid t,c) \cdot \log P(m \mid t,c)$$

To illustrate this relationship, consider three types of task representation units: (1) units necessary for task performance, (2) units that interfere with task performance, and (3) units that are irrelevant to the task. Suppose, for instance, that a unit h_1 must remain near 1 for the model to perform well, and ablating it (setting it to 0) reduces task performance to chance; in this case, the marginal probability $P(m_1 \mid t,c)$ would be near 1. Conversely, if a unit h_2 interferes with performance, its activation would reduce accuracy to chance, and $P(m_2 \mid t,c)$ would be near 0. Both h_1 and h_2 are thus causally relevant to the task and favor specific mask values (1 and 0 respectively), increasing the concentration of the ablation mask distribution. In contrast, an incidental unit h_3 that does not significantly affect outcomes will have $P(m_3 \mid t,c)$ near 0.5, decreasing concentration. Since each incidental unit contributes an independent binary choice (0 or 1) without affecting correctness, the number of high-probability masks grows exponentially with the number of incidental units, thereby increasing entropy. Thus, higher entropy reflects task representations with fewer causally relevant units and more incidental units.

Similarly, the entropy of an individual representational unit h_i over a specific task reflects how strongly the model depends on the particular unit's value. To compute this, we start with the marginal probability $P(m_i \mid t,c)$, which represents the likelihood of the unit h_i being active (not ablated) during successful task performance:

$$P(m_i \mid t,c) = \sum_{m} m_i \cdot P(m \mid t,c).$$

The marginal entropy $H(m_i \mid t,c)$, which is bounded between 0 and 1, is defined as

$$H = -p\log p - (1-p)\log(1-p).$$

where $p = P(m_i \mid t,c)$.

Unit importance and representational distributedness

The value of a representational unit is the atom of information processing in a neural system, and the collection of such units form distributed representations (Hinton, McClelland, & Rumelhart, 1986) that has been a rich topic of study in models both large (Arora, Li, Liang, Ma, & Risteski, 2018; Bricken et al., 2023) and small (Hinton, 1986). Although the representational unit value and a distributedness metric such as the L_1-norm of a task representation vector are easy to access and compute, these observational measures permit at most correlational interpretations and cannot provide causal interpretations. Here, we leverage the AMD to compute an analog of a representational unit value and representational distributedness with a causal interpretation, i.e. which units are actually relevant to the task.

We measure the model's reliance on the representational unit h_i using $1 - H(m_i \mid t,c)$ (where H is bounded between 0 and 1), which we refer to as unit importance. While this measure is correlated ($r = 0.661$) with the actual representational unit's value $h_i(t)$, its deviation indicates when the representational unit is not strictly necessary for the model to perform well. As shown in Figure 2, while unit importance generally increases with $h_i(t)$, a large number of representational units reflect low importance despite high activation values.

As might be expected, tasks with more distributed representations rely on a greater number of representational units with high importance, whereas more localized representations concentrate importance on only a few units. Thus, marginal entropy also provides a way to measure the effective representational distributedness of a task across the d representational units. We quantify this by summing across all d representational units: $d - \sum H(m_i \mid t,c)$. Naturally, this relates to the L_1-norm of the activation values ($\sum |h_i(t)|$, $r = 0.922$), but with

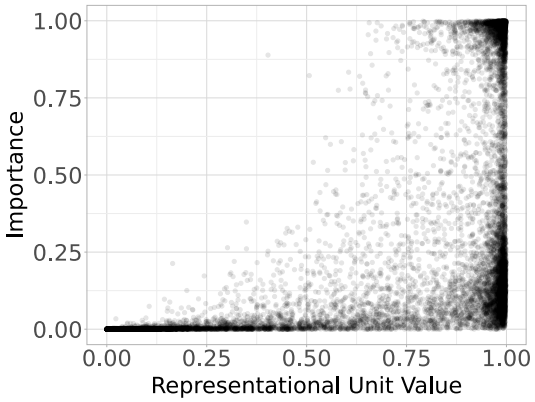

Figure 2: Task representation unit values $h_i(t)$ and their importance $1 - H(m_i \mid t, c)$.

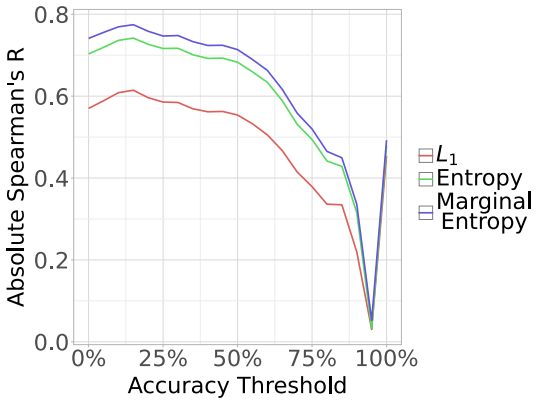

Figure 3: Correlation between and task representation metrics and task acquisition order along different accuracy thresholds. Task acquisition order is defined as the order that a model's task accuracy first exceeds the specified threshold.

the guarantee of a causal, rather than merely observational or correlational, interpretation.

Manifold complexity The distributed representations of neural networks often reflect dependencies, such that the activation of some units change how other units are used by the network by warping the representational manifolds (Fakhar & Hilgetag, 2022; Giallanza et al., 2024). We quantify this manifold complexity by measuring the joint entropy $H(m \mid t, c)$ of the AMD, which captures the full statistical dependencies between representational units, offering a more holistic view than the marginal entropy sum $\sum H(m_i \mid t, c)$. Comparing these two entropic measures allows us to quantify the information contained in higher-order dependencies, expressed as a normalized entropy drop:

$$\Delta H = 1 - \frac{H(m \mid t, c)}{\sum_i H(m_i \mid t, c)}.$$

The value of ΔH represents the proportion of entropy attributed to higher-order dependencies, providing a measure of the manifold complexity of task representations.

In the ISC model, we observe an average entropic reduction of $\Delta H = 4.62\%$, indicating that task representations are predominantly modular. This suggests minimal reliance on higher-order interdependencies among units, which may be expected for a simple feed-forward network designed for a specific set of semantic cognition tasks.

Task differentiation Progressive differentiation of internal representations is a fundamental property of neural networks that offers a unifying framework for developmental psychology and statistical learning (McClelland & Rogers, 2003; Munakata & McClelland, 2003). However, while it is possible to describe some of these phenomena through mathematical theory, existing tools are often limited to certain architectural constraints, such as containing purely linear layers (Saxe et al., 2019; Lampinen & Ganguli, 2018).

We hypothesized that the causal relevance and distributedness of task representation units relate to how the model differentiates representations during training. Specifically, the null-task—represented by a zero-vector input and output—should drive early-learned tasks to push unit values away from zero. In contrast, later-learned tasks may rely on fewer units by incrementally diverging from existing representations.

To test this hypothesis, we recorded the order in which the accuracy first rose above 0% for each task, which corresponds to the initial accuracy for all tasks due to the presence of the null-task. We then compared this rank metric to the two entropic measures and the L_1-norm of each task representation using Spearman's rank-order correlation. Consistent with our hypothesis, we found a high absolute correlation between the order that the model learned each task and the two entropy measures ($r = 0.708$ for joint entropy, $r = 0.746$ for marginal entropy). Despite its relatively high correlation with entropy, the L_1-norm had only a moderate correlation of $r = 0.573$ with the task acquisition order. This indicates that the ablation masks' capacity to distinguish between causally relevant and merely incidental representational values may allow them to capture model properties more accurately. Moreover, when we compared the correlation using various accuracy thresholds as shown in Figure 3, we found that the absolute correlation begins to drop around an accuracy threshold of 15%, suggesting that the entropic measures are sensitive to how the model initially allocates representational units but less to how it refines their values during training.

Mutual information, reverse inference, and polysemanticity

We now consider how the AMD can be used to address the reverse inference problem of identifying the task from the representational unit activations. That is, beyond evaluating whether a representational unit h_i contains sufficient information to decode the task, we ask whether h_i is *causally involved* in encoding the task, i.e. whether the unit is mono- or polysemantic (Arora et al., 2018; Elhage et al., 2022; Bricken et al.,

2023) with respect to the various tasks in the dataset.

By leveraging the AMD, we isolate h_i's necessity for task performance, distinguishing incidental activations from task-relevant contributions and quantifying how its influence is distributed across multiple tasks. This conditional probability is given by:

$$P(t \mid m, c) = \frac{P(m \mid t, c) \cdot P(t \mid c)}{\sum_{t'} P(m \mid t', c) \cdot P(t' \mid c)}$$

Marginalizing over individual units gives the probability of a task given the activation state of a unit:

$$P(t \mid m_i, c) = \frac{\sum_{m'} m'_i \cdot P(t \mid m', c) \cdot P(m', c)}{\sum_{t', m'} m'_i \cdot P(t' \mid m', c) \cdot P(m', c)}$$

where m'_i is the value of the i^{th} bit in mask m'.

Using the conditional task distribution, we compute the mutual information by measuring the reduction in entropy, which we normalize for interpretability:

$$I_n(t, m \mid c) = 1 - \frac{H(t \mid m, c)}{H(t \mid c)}, \quad I_n(t, m_i \mid c) = 1 - \frac{H(t \mid m_i, c)}{H(t \mid c)}$$

The entropy of the task distribution conditioned on unit activation provides a measure of task polysemanticity. For example, consider a representational unit that encodes exactly one task t', so that $P(t' \mid m_i, c) = 1$ and 0 for all other tasks. In this case, the resulting entropy $H(t' \mid m_i, c)$ and the normalized mutual information $I_n(t', m_i \mid c)$ are 0 and 1 respectively. Conversely, a unit that provides no task information when considered independently, so that $P(t \mid m_i, c) = P(t)$, will result in $I_n(t, m_i \mid c) = 0$. Thus, I_n provides a bounded measure of a unit's task specificity, ranging from 0 to 1. While $I_n = 1$ indicates perfect determinism and $I_n = 0$ indicates no task relevance, any $I_n > 0$ indicates that a unit encodes some useful information–just not in isolation. Instead, its contribution must be combined with others, either conjunctively or compositionally, to represent the full task structure. Thus, this measure reflects polysemanticity with respect to tasks rather than sub-task concepts, as a unit with low I_n for tasks may still encode meaningful sub-task structure, e.g. semantic features across multiple tasks, allowing it to participate in structured, compositional representations that support task generalization.

We compare the difference between the two measures computed on the full mask distributions and on the marginal unit distributions. As shown in Figure 4, we find that individual units share very little mutual information with tasks when considered independently, reducing entropy by about 4.21% in most cases. In contrast, ablation masks are significantly more informative, reducing uncertainty by an average of 82.6%. This suggests that the representational units individually encode little information about particular tasks, and that the model relies on ensembles of representational units – that is, distributed representations. At the same time, because individual units participate in multiple task representations, this also implies a form of polysemanticity, where units contribute to multiple tasks in a structured way rather than encoding distinct tasks independently.

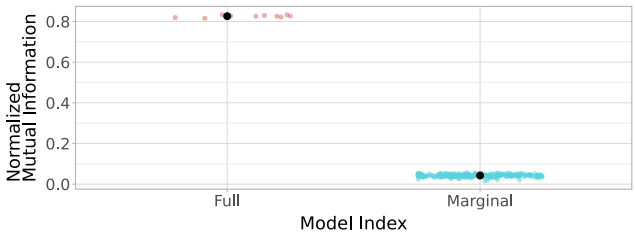

Figure 4: Percentage of normalized mutual information captured by the full AMD, $P(t \mid m)$, and by marginal unit distributions, $P(t \mid m_i)$. Each point represents a different model seed in 'Full' (10 total) or a combination of model seeds and representational units (240 total) in 'Marginal'.

Task similarity

Thus far, we have focused on measures targeted at understanding individual task representations. In this section, we extend our analysis to compare the similarity between task representations. Because the ablation mask distributions reflect causal relevance and higher-order dependencies between representational units, they capture relationships that vector-based measures, such as cosine or Euclidean distance, may overlook. For example, under cosine or Euclidean distance, the vector [1,1] would be considered further from [0,0] than [0,1], even if the second dimension is not meaningfully used by the model. To compare the task similarities using the ablation mask distributions, we turn to two distance measures well-suited for comparing probability distributions: KL-divergence and Wasserstein distance.

The KL-divergence $D_{KL}(P\|Q)$ measures the difference between two probability distributions by quantifying the information lost when approximating one distribution (Q) with another (P):

$$D_{\text{KL}}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

However, because $D_{KL}(P\|Q)$ is inherently asymmetric, we use the symmetrized KL-divergence $D^S_{KL}(P\|Q)$ which combines $D_{KL}(P\|Q)$ and $D_{KL}(Q\|P)$ into a bidirectional measure:

$$D^S_{\text{KL}}(P\|Q) = \frac{1}{2} D_{\text{KL}}(P\|Q) + \frac{1}{2} D_{\text{KL}}(Q\|P).$$

We also consider the Wasserstein distance $W(P,Q)$, which quantifies the minimal cost of transforming one distribution into another:

$$W(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[d(x,y)],$$

where $\Gamma(P,Q)$ is the set of joint distributions (couplings) with marginals P and Q, and $d(x,y)$ is the Hamming distance between mask configurations x and y, i.e., the number of 1's or 0's that need to be flipped to transform one mask into another.

A key difference between KL-divergence and Wasserstein distance is that the latter is informed by the distance metric

(Hamming distance) while KL-divergence is not. For example, if two distributions agree on all but two masks, the KL-divergence between them will depend only on the differing amounts of mass placed on these masks. Wasserstein distance is sensitive to this difference in mass, and also to the Hamming distance between the masks. In particular, the Wasserstein distance will be greater if the two masks share fewer bits in common, whereas the KL-divergence is not sensitive to this.

To contextualize these probabilistic measures, we compare them to more conventional vector-based metrics: cosine similarity and Euclidean distance. While cosine similarity and Euclidean distance do not account for higher-order dependencies or causal relevance, they are reasonable points of comparison as they are used widely in assessing the similarity of vector-based representations, both in neural networks and empirical neural data (Kriegeskorte, Mur, & Bandettini, 2008). Furthermore, they are useful in the evaluating representations in the ISC model, given the relative simplicity of its representational manifold as evidenced by the low entropy drop ΔH. Additionally, we introduce a non-parametric measure of task similarity that we refer to as mask-performance correlation (MPC), which compares the correlation between the accuracies of two tasks when the same mask is applied. This measure provides a direct link between ablation masks and task performance without incorporating the importance weighting involved in probabilistic measures computed over the posterior distribution. Specifically, MPC measures correlation using $P(c|t,m)$ without further weighting, whereas the AMD metrics weight the distances by $P(m|c,t)$

To compare the various measures, we conduct a representational similarity analysis (RSA) (Kriegeskorte et al., 2008), computing the absolute Spearman correlation between metrics across task pairs (Figure 5). KL-divergence and Wasserstein distance exhibit strong correlation ($r = 0.73$), highlighting their shared reliance on posterior mask distributions. The especially high correlation ($r = 0.89$) between Wasserstein distance and cosine similarity and suggests that the probabilistic framework preserves much of the structural information captured by traditional similarity metrics. Both measures align with intuitive similarities between certain tasks, such as between 'lexical expressions' and 'synonyms', or between 'related actions' and 'external features'. In contrast, MPC exhibits a relatively weak correlation with other measures. For instance, its correlation with cosine similarity drops to $r = 0.50$, suggesting that posterior weighting is important for preserving representational fidelity, beyond merely considering the outcome of performance for each task.

Discussion

In this paper, we have introduced a novel probabilistic framework for studying task representational structure in neural networks. Unlike simple ablation which only evaluates downstream effects on task performance, our approach uses a Bayesian perspective that reconstructs task representations as posterior distributions over ablation masks, allowing for causal interpretation of task representations. This probabilistic approach facilitates the use of tools from information theory and optimal transport, enabling a deeper exploration of task representations that is sensitive to the structure of the representational manifold. For example, measures such as entropy and mutual information can be used to quantify how a neural networks distributes information in complex manifolds.

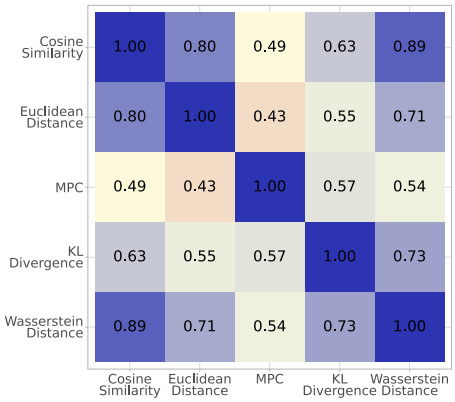

Figure 5: Spearman correlation between similarity measures.

Our framework has several limitations that prompt further research. First, although we introduce metrics to quantify representational phenomena (e.g., manifold complexity and statistical dependence), these abstract constructs are hard to validate and require additional theoretical and empirical work to connect with observable phenomena in neural networks and cognitive systems. Second, our analyses focus on a single dataset and model—the ISC model trained on the Leuven Concepts Database. While this choice offers strong psychological relevance and interpretability in a controlled setting, it leaves room to explore more complex architectures and diverse datasets for additional insights into task representations and cross-domain generality. Finally, scaling the Bayesian approach is difficult; computing the posterior over ablation masks becomes increasingly hard as the number of representational units grows, necessitating careful approximations for large-scale models. Promising directions include generative neural networks and sampling-based methods like Markov Chain Monte Carlo (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), though applying to foundational models with billions of parameters may still present a challenge.

In conclusion, this work introduces a probabilistic framework for understanding task representations in neural networks, providing a principled approach to uncover causal relationships and representational complexity. While further development and scaling are needed, our approach lays a foundation for future research into task representations across both natural and artificial systems. We hope this framework inspires new insights into the principles governing learning and cognition in neural network-based architectures.

Acknowledgments

We thank Tyler Giallanza for providing materials related to the ISC model, Legasse Remon for assistance with running experiments, and Arjun Menon for helping organize the item property taxonomy table.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, *6*, 483–495.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... others (2024). Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., ... Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. (https://transformer-circuits.pub/2023/monosemantic-features/index.html)

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, *97*(3), 332.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 dutch words in a continuous task. *Behavior research methods*, *40*(1), 198–205.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., ... Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread*.

Fakhar, K., & Hilgetag, C. C. (2022). Systematic perturbation of an artificial neural network: A step towards quantifying causal contributions in the brain. *PLOS Computational Biology*, *18*(6), e1010250.

Giallanza, T., Campbell, D., Cohen, J. D., & Rogers, T. T. (2024). An integrated model of semantics and control. *Psychological Review*.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 8).

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, volume 1: Foundations* (pp. 77–109). Cambridge, MA: MIT Press.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., ... Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, *3*(7), pgae233.

Lampinen, A. K., & Ganguli, S. (2018). An analytic theory of generalization dynamics and transfer learning in deep linear networks. *arXiv preprint arXiv:1809.10374*.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, *4*(4), 310–322.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*(4), 413–429.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*. (https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html)

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*(45-76), 26.

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 506–515.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Storms, G. (2001). Flemish category norms for exemplars of 39 categories: A replication of the battig and montague (1969) category norms: Pet studies. *Brain*, *124*, 1619–1634.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Wu, L.-l., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica*, *132*(2), 173–189.

Yu, D., Kaur, S., Gupta, A., Brown-Cohen, J., Goyal, A., & Arora, S. (2023). Skill-mix: A flexible and expandable family of evaluations for ai models. *arXiv preprint arXiv:2310.17567*.