TASK-MEDIATED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditionally, unsupervised representation learning is used to discover underlying regularities from raw sensory data without relying on labeled data. A great number of algorithms in this field resorts to utilizing proxy objectives to facilitate learning. Further, learning how to act upon these regularities is left to a separate algorithm. Neural encoding in biological systems, on the other hand, is optimized to represent behaviorally relevant features of the environment in order to make inferences that guide successful behavior. Evidence suggests that neural encoding in biological systems is shaped by such behavioral objectives. In our work, we propose a model of inference-driven representation learning. Rather than following some auxiliary, a priori objective (e.g. minimization of reconstruction error, maximization of the fidelity of a generative model, etc.) and indiscriminately encoding information present in an observation, our model learns to build representations that support accurate inferences. Given a set of observations, our model encodes underlying regularities that de facto are necessary to solve the inference problem in hand. Rather than labeling the observations and learning representations that portray corresponding labels or learning representation in a self-supervised manner and learning explicit features of the input observations, we propose a model that learns representations that implicitly shaped by the goal of correct inference.

1 INTRODUCTION

One of the central problems in unsupervised representation learning traditionally has been to find a way of learning useful representations when target values are not available or simply do not exist (Hinton & Sejnowski, 1999; LeCun et al., 2015). A common approach to this problem is to essentially compress high dimensional input s.t. it can be later reproduced from a low dimensional representation with minimum deviation from either the source or any other proxy objective. For instance, autoencoders learn latent representations by minimizing the reconstruction error (Hinton & Salakhutdinov, 2006); variational autoencoders additionally constrain the representation which allows for learning the factors of variation (Kingma & Welling, 2013; Higgins et al., 2017a); generative models can learn representations by maximizing the fidelity of generated outputs (Donahue et al., 2016). Once representations are learned, their utility is usually evaluated on a separate downstream task (Hsu et al., 2018; Higgins et al., 2017b; Metz et al., 2019; Laversanne-Finot et al., 2018). On the other hand, there is a significant amount of evidence that suggests that perceptual neural encoding in biological systems is shaped by the process of inference supporting effective behavior (Janzen & Turennout, 2004). Theory of embodied cognition suggests that the process of information internalization requires physical interaction with the environment (Calvo & Gomila, 2008). In our work, we propose one of the possible ways of learning representations of abstract concepts from unlabeled data that in turn support the inference. We argue that an abstract surrogate objective such as correct inference not only yields adequate representations but also a biologically valid way of learning representations.

We organized the paper by first defining and describing our dataset and the model in Section 2 and Section 3 respectively. In Section 4, we describe the results and provide an analysis of the network behavior and performance, demonstrating the robustness of learned representations. Then, in order to contextualize our work, we provide a summary of related research and empirical evidence that supports our approach.

2 DATASET

In our work, we use a synthetic dataset that consists of several types of inference tasks and observations O. In general, the task is to make accurate inference based on a problem P which requires correct interpretation of given observations O. Every observation is generated such that it follows a certain pattern or regularity (i.e. regularities are implicitly encoded in the observations). Based on the regularities encoded in the tasks, they can be divided into three categories: (1) binary operations, (2) contextual 2-armed bandit, and (3) associative recall. In the case of binary operations, a single observation can be often ambiguous hence we use a set of three observations to reduce ambiguity. Finally, we use the results of the inference problems as targets. Learning algorithm does not use labels of the regularities placed in the observations. The data is presented in the form of tensors of binary activations. For the simplicity of data interpretation, we address each element as if it was a number.

X			y
5,	2,	11	
13,	27,	110	
170,	28,	45	
13,	27,	_	110

X			y
5,	2,	1	
5,	2,	0	
2,	5,	1	
5,	2,	_	1

Figure 1: (Top) Structure and possible interpretation of a training entry. Here, the underlying regularity is based on binary function $f(x_1, x_2)$ is: $x_3 = x_1 + x_2$. (Bottom left) Structure of Associative Recall task: x_1 and x_2 together constitute a key and x_3 is the target value. (Bottom right) Structure of Contextual 2-Armed Bandit task: $x_3 = 0$ means that the left arm can be associated with the reward, if $x_3 = 1$, then it is the right arm (here, the arm 13 should be preferred.)

2.1 BINARY OPERATIONS

This part of the dataset includes 7 functions:

1. $x_3 = x_1 + x_2 + c$, where $c \in [0, 1, 2, 3]$
2. $x_3 = \mathrm{sd}(x_1) + \mathrm{sd}(x_2)$, where sd is digit sum of an integer
3. $x_3 = x_1 > x_2, x_3 \in 0, 1$
4. $x_3 = x_1 < x_2, x_3 \in 0, 1$
5. $x_3 = x_1 - x_2, s.t.\ x_3 > 0$
6. $x_3 = x_2 - x_1, s.t.\ x_3 > 0$
7. $x_3 = x_2 \operatorname{div} x_1$, where div is integer division

Each of the observations is always structured such that the third number in it is a result of a mathematical or logic binary operation applied to the first two numbers: $\boldsymbol{x}_3 = f(\boldsymbol{x}_1, \boldsymbol{x}_2)$. To make an inference, the model needs to determine the function f based on the observations $\textsf{\textbf{O}}$ and apply it to the given problem $\textsf{\textbf{P}}$: $\hat{\boldsymbol{y}} = f(\textsf{\textbf{P}})$. Within each function included in the dataset, each pair of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is unique.

2.2 Contextual 2-Armed Bandit

The problem of the contextual 2-armed bandit is a variant of a well-known multi-armed bandit problem (Auer et al., 2002). The 2-armed bandit problem is usually used in meta-learning research. We, however, reframe it as an inference problem. Given a set of observations, the task is to pick either left or right arm. Each arm has a number associated with it. In the contextual bandit problem, it is not necessarily the left or the right arm that is more likely to score but the one with a label (number in our dataset) that, in the observations, corresponds with the right decision. The goal is therefore to pick the arm which is more likely to score. In our dataset we use the following reward distribution: $66\frac{2}{3}\%$ or $33\frac{1}{3}\%$.

2.3 Associative Recall

Associative recall task is reminiscent of a simple short-term memory task: given a dictionary (i.e. a set of (key, value) tuples), the task is to recall the value that corresponds to the given key in the dictionary. Every key in the dataset is unique, hence, every association is also a unique combination.

3 Model

3.1 Model Architecture

Our model is a two-tailed neural network. It takes a set of observations $\textsf{\textbf{O}}$ and inference problem $\textsf{\textbf{P}}$ and makes an inference $\hat{\boldsymbol{y}} : \hat{\boldsymbol{y}} = H(\textsf{\textbf{O}}', \textsf{\textbf{P}}'; \theta^H)$. The first tail takes observations one-by-one, consolidates them in the internal state of a recurrent cell, passes combined representation through a bottleneck $\boldsymbol{r}$, and finally combines its output $\textsf{\textbf{O}}' = F(\textsf{\textbf{O}}; \theta^F)$ with the output of the second tail $\textsf{\textbf{P}}' = G(\textsf{\textbf{P}}; \theta^G)$. Second tail and the head components of the network are stacks of fully connected layers.

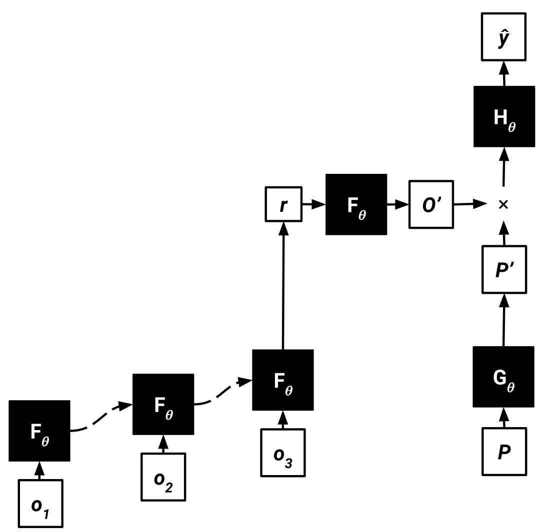


Figure 2: Model architecture

3.2 Learning Strategy

One of the major goals of this work is to introduce a representation learning mechanism that combines several observations and is focused on discovering regularities underlying them (implicit information) rather than on compressing input data into more interpretable features of explicitly presented information. We suggest that one of the possible ways to do that is to learn representations not independently but in the process of observation facilitated inference. Hence, our dataset only contains the results for the inference problems but does not label the underlying regularities. In fact, in some cases, it is possible to assign several labels to a single regularity (for details see Section 4) which makes labeling counterproductive. Such an approach has its own limitations, yet it is based on certain aspects of representation learning in biological systems (see Section 5.2 for details).

3.2.1 OBJECTIVE FUNCTION

Our model is trained to minimize the discrepancy between inferred result and actual result. We also use a regularization term R to make representation r more efficient, as suggested in (Kingma & Welling, 2013; Higgins et al., 2017a). The loss function is the following:

$$L = \sum_{i=0}^{n} (\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)^2 + \beta * R \tag{1}$$

We use multivariate KullbackLeibler divergence (D_{KL}) (2) between a diagonal multivariate normal, and a standard normal distribution (Kullback and Leibler, 1951) for representation regularization purposes:

$$R = D_{KL}(\mathcal{N}(\mu, \sigma^2) \,||\, \mathcal{N}(\mathbf{O}, \mathbf{I})) = \frac{1}{2} \sum_{i=0}^{m} \sigma_i^2 + \mu_i^2 - \ln \sigma_i^2 - 1 \tag{2}$$

Thus, the objective is the following:

$$\arg\min_{\theta} L \tag{3}$$

3.3 IMPLEMENTATION

We implemented our model in Tensorflow (Abadi et al., 2016) and used the Adam optimization algorithm for training (Kingma & Ba, 2014). We use a synthetic dataset generated according to Section 2. The size of the dataset depends on a set of variables and can be customized. The content of the dataset can also be customized by excluding and including certain problems. Our model was trained on all problems described in Section 2 together. We split the dataset 50/50 for training and testing respectively. In this paper, we report only test performance. Our model utilizes 1,277,084 trainable parameters θ. Input size is 192: observations O take (i.e. 48 each observation $\boldsymbol{o}_i$), and 48 for the problem $\mathsf{\mathbf{P}}$. The number of observations is a variable. In this work, we use 3 observations $\boldsymbol{o}_i$, yet our model shows adequate results for both fewer and more observations. For recurrent layer we use Gated Recurrent Unit cell (Cho et al., 2014) with an internal size of 96. The size of the bottleneck $\boldsymbol{r}$ is 8.

4 RESULTS

In our work, we demonstrate a model of representation learning mediated by an inference process. We show that certain abstract concepts can be learned from observations only retrospectively in the process of making inferences. Empirical evidence suggests that, when it comes to representing abstract concepts, biological systems, rather than representing a maximum number of features, facilitate successful behavior with the selective neural encoding of features relevant at the behavioral level. Since it is difficult to imagine an a priori rule that can extract relevant information, we trained our model using the objective that optimizes for the accuracy of the inference. Overall test performance of our model (98.6±0.1%) suggests that it was able to learn representations that facilitate inference.

To visualize the representations learned by our model we reduced their dimensionality to 2 via an autoencoder and plotted the results. We also plotted the centers of the manifolds of representations and

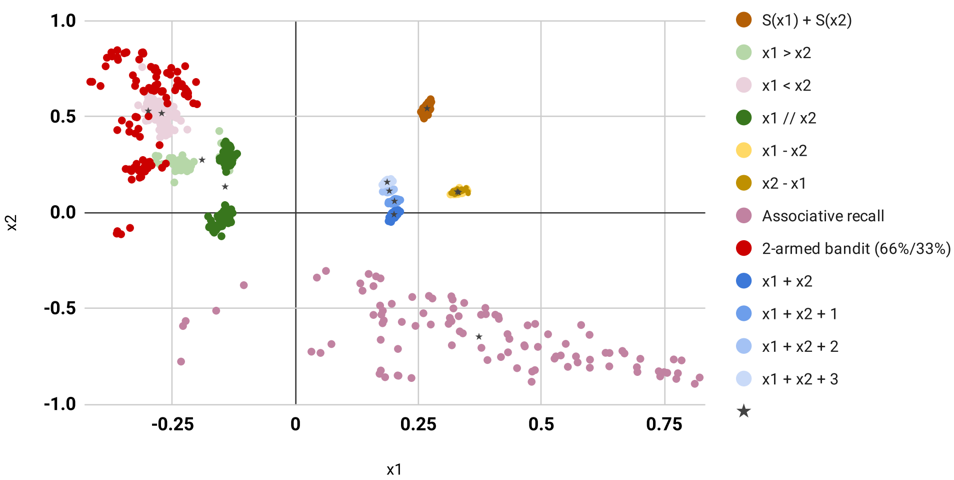


Figure 3: Learned representations of mathematical problems, associative recall, and contextual 2-armed bandit

calculated the distances between these points to illustrate the relative position of the representations. Figures 2 and 3 suggest that our model learned certain relationships between the concepts that these representations depict. For instance, virtually complete overlap between representations of $\boldsymbol{x}_3 = \boldsymbol{x}_1 - \boldsymbol{x}_2$ and $\boldsymbol{x}_3 = \boldsymbol{x}_2 - \boldsymbol{x}_1$, s.t. $\boldsymbol{x}_3 > 0$ shows that our model found that these operations are equivalents of the absolute difference between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Close proximity and order in the group of representations for $\boldsymbol{x}_3 = \boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{c}$, where $\boldsymbol{c} \in [0, 3]$ shows that the model discovered the relationship between these operations. Overlap between special cases of $\boldsymbol{x}_3 = \boldsymbol{x}_1 > \boldsymbol{x}_2$ and $\boldsymbol{x}_3 = \boldsymbol{x}_1 \operatorname{div} \boldsymbol{x}_2$ can be interpreted as well: for all $\boldsymbol{x}_1$ and $\boldsymbol{x}_2 : \boldsymbol{x}_1 > \boldsymbol{x}_2$ both functions will output 0.

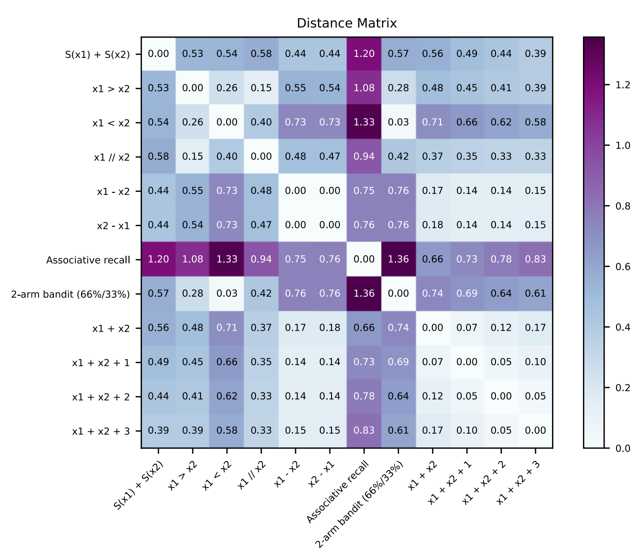
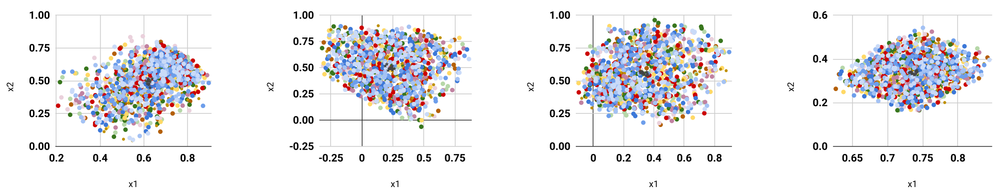


Figure 4: Distance between manifolds of learned representations

To prove that our model learns representations empirically shaped by the objective of the inference as opposed to an a priori meaning of underlying regularities, we coupled observations of binary operations with unrelated inference problems (e.g. $\mathbf{O}_{\boldsymbol{x}_3=\boldsymbol{x}_1>\boldsymbol{x}_2}$ with $\mathbf{P}_{\boldsymbol{x}_3=\boldsymbol{x}_1 \operatorname{div} \boldsymbol{x}_2}$ for all $\mathbf{O}_{\boldsymbol{x}_3=\boldsymbol{x}_1>\boldsymbol{x}_2}$ and $\mathbf{P}_{\boldsymbol{x}_3=x_1 \operatorname{div} \boldsymbol{x}_2}$). Performance of the model did not change significantly. Therefore, we concluded that, unless observation contains necessary information (e.g. like in the case of associative recall or

2-armed bandit tasks), a correct inference can be supported by any regularity that empirically appears to be useful.

We compared performance of representation learning component of our model with the performance of β-VAE (with $\beta \in [0, 1, 16, 64]$). The structure of encoder and decoder networks in the β-VAE was precisely the same as of the representation learning component of our model. Neither AE (i.e. $\beta = 0$), nor VAE ($\beta = 1$), nor β-VAE ($\beta \in [16, 64]$) managed to learn meaningful representations of the observations presented in our dataset. Given the same amount of training, baseline models did not converge to a reasonable error rate.



(a) $\beta = 64$ (b) $\beta = 16$ (c) $\beta = 4$ (d) $\beta = 1$

Figure 5: Representation learned by β-VAE (with $\beta \in [1, 16, 64]$)

5 Related Work

5.1 Representation Learning

One of the major goals of representation learning has been an ability to encode raw data in the way such an encoding could be useful in subsequent learning of a downstream task (Kingma & Ba, 2014; Rasmus et al., 2015; Hsu et al., 2018) or have useful qualities, such as interpretability, smoothness of the manifold, explanatory power, sparsity, disentanglement of the underlying factors of variation, etc. (Kingma & Welling, 2013; Higgins et al., 2017a; Donahue et al., 2016; Bengio et al., 2014).

Hinton & Salakhutdinov (2006) showed that autoencoders (AE) can transform high dimensional data into low dimensional representations that minimize reconstruction error. Kingma & Welling (2013) presented Variational Auto Encoders (VAE) an approach to learning representations that encode high-level factors of variation. More recent, Higgins et al. (2017b) pre-trained β-VAE to encode basic visual concepts of the visual environment, effectively disentangling such variational factors. In turn, latent representations produced by β-VAE supported a reinforcement learning model and have proven to be particularly useful for transfer learning. Similarly, Laversanne-Finot et al. (2018) utilized β-VAE to learn representations that contain independent features. Then, a separate model used pre-trained β-VAE to support curiosity-driven exploration in a robotic arm task. Hsu et al. (2018) used several unsupervised embedding algorithms to learn representations useful in subsequent object discrimination procedure. Their results suggest that the use of embeddings improved the performance of the discrimination model.

Goodfellow et al. (2014) demonstrated the ability of Generative Adversarial Networks (GANs) to learn a mapping from latent distribution to data distribution and captures semantic variation features. Further work, Bidirectional Generative Adversarial Networks (BiGAN) (Donahue et al., 2016), made it possible to learn a mapping from data to latent representations useful for a supervised discrimination task.

5.2 Task-Mediated Neural Encoding

A significant body of empirical evidence and theoretical work suggests that neuronal encoding emerges and further develops mediated by the behavioral objectives (Oxenham (ed.) & Oxenham, 2005; Kuchibhotla & Bathellier, 2018; Kraus et al., 2014; Janzen & Turennout, 2004). Biological systems, in order to succeed in achieving these objectives, need to make inferences based on sensory input. In turn, the efficiency of the inference is contingent on the ability to encode relevant information. Sustained activity in PFC support context-relevant representations when a task requires different behavior based on the context (Rikhye et al., 2018; Bolkan et al., 2018). The visual system of a fly is tuned to process and represent optic flow, which is perhaps the most important visual feature at the behavioral level of a relatively simple animal (Egelhaaf et al., 2002). Recent research suggests that musical training improves auditory processing by enhancing neural encoding of meaningful acoustic features which, in turn, benefits language and cognition (Kraus et al., 2014; Tierney et al., 2013; Parbery-Clark et al., 2009; Kraus et al., 2010). Implicit relevance of objects placed at key points was associated with neural activity in the parahippocampal gyrus (area of the brain implicated in visual navigation) whereas explicit relevance improved object recognition processing in the visual cortex (Janzen & Turennout, 2004). Similarly, in auditory cortex, encoding has been shown to be mediated by ascending neuromodulation and top-down attention (Kuchibhotla & Bathellier, 2018; Caras & Sanes, 2017).

6 Discussion

The ability to perceive regularities in the raw perceptual data one the most important skills that biological systems develop very early and continue to develop and which facilitates learning and other cognitive abilities. Development of this ability does not require explicit training as it is observed in various biological systems and perhaps one of the fundamental mechanisms in learning. Likewise, unsupervised learning, a discipline in machine learning, has been focused on problems that cannot and do not have explicit training or labels. We showed that the absence of labels does not automatically mean that the process of learning of the regularities is driven by a single universal objective (e.g. minimization of reconstruction error). Rather, learning of representations of relevant information can be shaped by behavioral objectives and enables learning of abstract concepts that cannot be learned solely from observations.

The utility of the representations learned in such a manner will depend on their behavioral relevance and experience. We showed that simple model that is given a reasonable objective, can learn representations that even very clever encoding algorithms could not in a self-supervised manner. Hence, in conclusion, we would like to emphasize the role of intrinsic motivation (i.e. objective) and the environment in learning useful representations.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016. URL `http://arxiv.org/abs/1603.04467`.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL `https://doi.org/10.1023/A:1013689704352`.

Yoshua Bengio, É ric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–226–II–234. JMLR.org, 2014. URL `http://dl.acm.org/citation.cfm?id=3044805.3044918`.

Scott S. Bolkan, Joseph M. Stujenske, Sebastien Parnaudeau, Timothy J. Spellman, Caroline Rauffenbart, Atheir I. Abbas, Alexander Z. Harris, Joshua A. Gordon, and Christoph Kellendonk. Thalamic projections sustain prefrontal activity during working memory maintenance. *Nature Neuroscience*, 20(7):987–996, 2018. doi: 10.1038/nn.4568. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501395/`.

Paco Calvo and Toni Gomila. *Handbook of Cognitive Science: An Embodied Approach*. Perspectives on Cognitive Science. Elsevier Science, 2008. ISBN 9780080466163. URL `https://books.google.com/books?id=L9qFD0TMeT4C`.

Melissa L. Caras and Dan H. Sanes. Top-down modulation of sensory cortex gates perceptual learning. *Proceedings of the National Academy of Sciences*, 114(37):9972–9977, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1712305114. URL https://www.pnas.org/content/114/37/9972.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016. URL http://arxiv.org/abs/1605.09782.

Martin Egelhaaf, Roland Kern, Holger Krapp, Jutta Kretzberg, Rafael Kurtz, and Anne-Kathrin Warzecha. Neural encoding of behaviourally relevant visual-motion information in the fly. *Trends in neurosciences*, 25:96–102, 03 2002. doi: 10.1016/S0166-2236(02)02063-5.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017a. URL https://openreview.net/forum?id=Sy2fzU9gl.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning, 2017b.

Geoffrey E. Hinton and Russ R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504507, 2006.

Geoffrey E. Hinton and Terrence J. Sejnowski. *Unsupervised Learning: Foundations of Neural Computation*, volume 1. MIT Press, 1999.

Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. *CoRR*, abs/1810.02334, 2018. URL http://arxiv.org/abs/1810.02334.

Gabriele Janzen and Miranda van Turennout. Selective neural representation of objects relevant for navigation. *Nature Neuroscience*, 7(6):673–677, 2004.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

Nina Kraus, Erika Skoe, Alexandra Parbery-Clark, and Richard Ashley. Experience-induced malleability in neural encoding of pitch, timbre, and timing. implications for language and music. *Annals of the New York Academy of Sciences*, 1169:543557, 2010. doi: 10.1111/j.1749-6632.2009.04549.x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2810198/.

Nina Kraus, Jessica Slater, Elaine C. Thompson, Jane Hornickel, Dana L. Strait, Trent Nicol, and Travis White-Schwoch. Music enrichment programs improve the neural encoding of speech in at-risk children. *Journal of Neuroscience*, 34(36):11913–11918, 2014. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1881-14.2014. URL https://www.jneurosci.org/content/34/36/11913.

Kishore Kuchibhotla and Brice Bathellier. Neural encoding of sensory and behavioral complexity in the auditory cortex. *Current Opinion in Neurobiology*, 52:65–71, 10 2018. ISSN 0959-4388. doi: 10.1016/j.conb.2018.04.002.

Adrien Laversanne-Finot, Alexandre Péré, and Pierre-Yves Oudeyer. Curiosity driven exploration of learned disentangled goal spaces. *CoRR*, abs/1807.01521, 2018. URL http://arxiv.org/abs/1807.01521.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015. ISSN 0028-0836. doi: 10.1038/nature14539.

Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. Meta-learning update rules for unsupervised representation learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=HkNDsiC9KQ.

Andrew J. Oxenham (ed.) and Andrew J Oxenham. *Pitch: Neural Coding and Perception*. Springer, 2005.

Alexandra Parbery-Clark, Erika Skoe, Carrie Lam, and Nina Kraus. Musician enhancement for speech-in-noise. *Ear and Hearing*, 30(6):53–661, 2009. ISSN 0196-0202. doi: 10.1097/AUD.0b013e3181b412e9. URL https://insights.ovid.com/crossref?an=00003446-200912000-00002.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3546–3554. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5947-semi-supervised-learning-with-ladder-networks.pdf.

Rajeev V. Rikhye, Aditya Gilra, and Michael M. Halassa. Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, 21:1753–1763, 2018. doi: 10.1038/s41593-018-0269-z. URL https://doi.org/10.1038/s41593-018-0269-z.

Adam Tierney, Jennifer Krizman, Erika Skoe, Kathleen Johnston, and Nina Kraus. High school music classes enhance the neural processing of speech. *Frontiers in Psychology*, 4:855, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00855. URL https://www.frontiersin.org/article/10.3389/fpsyg.2013.00855.