ENTROPY CONSTRAINED INFORMATION BOTTLENECK

Lingyu Zhang

ABSTRACT

The information bottleneck (**IB**) is a principle for learning compressive representations for predictive tasks. While it has been controversial whether neural networks are inherently compressing features, studies using neural networks to explicitly parameterize the IB objective showed promising results in terms of generalization and robustness. The main challenge lies in the estimation and optimization of mutual information for high dimensional data. Specifically, when the mapping from input X to the encoded representation T is injective, the mutual information is either infinite or piece-wise constant and independent of model parameters. In this work we propose to tackle these issues by using deterministic encoding along with actual quantization on latents, rendering the problem a source compression. By doing so, finite non-trivial mutual information can be estimated. We further integrate variational entropy estimation commonly used in neural lossy compression to jointly optimize the encoder and entropy model. We evaluate our method on supervised and self-supervised learning tasks for generalization, and against adversarial attack for robustness. We show that our method is able to learn compact and sparse representations that exhibit improved generalization and stronger robustness against adversarial attacks. Code is available `https://github.com/lingyu98/cSCL`

1 INTRODUCTION

Formulating principles for learning optimal representations has been one of the fundamental problems in machine learning. The information bottleneck (**IB**) Tishby et al. (2000) introduced a principle for representation learning that trades off informativeness of the target variable and compressiveness with respect to the input variable, enforcing representations to be the minimal sufficient statistics of the input. The intuition is that when learning a predictive task $X \rightarrow Y$, the intermediate representation T should contain sufficient information about Y, but also minimal information about X, compressing out the information in X that is irrelevant of the task. The measurement for informativeness is the mutual information:

$$T^* = \arg\max_T I(Y;T) \quad \text{s.t.} \quad I(X;T) \leq r \tag{1}$$

Where the MI between the input variable and the representation $I(X;T)$ is bounded by some r. The above objective can be formulated as the following Lagrangian to circumvent the non-linear constraint:

$$T^* = \arg\max_T I(Y;T) - \beta I(X;T) \tag{2}$$

With β being the Lagrangian multiplier. Optimizing the objective above for different βs correspond to different points on the IB curve.

In Shwartz-Ziv & Tishby (2017), an attempt was made at explaining the generalization abilities of neural networks with the IB. The authors argue that neural networks learned with SGD is inherently optimizing for something approximating the IB objective. They identified in the training procedure a fitting phase that increased $I(X;T)$ and $I(T;Y)$ and a compression phase that decreased $I(X;T)$. The claim was controversial as Saxe et al. (2018) argue that the observed phenomenon was a result of saturating non-linearities and they questioned the causality between compression and generalization. More recently, Lorenzen et al. (2022) aimed at resolving the controversy by computing exact mutual information of quantized neural networks, confirming the existance of a compression phase and also its dependency on activation functions.

On the other hand, another line of research have been conducted in directly parameterizing the IB objective. The idea is appealing because many issues in deep learning could in principle be addressed by the IB. Over-fitting, spurious correlations, adversarial susceptibility, all can be seen as a consequence of learning irrelevant information. However, the authors of Tishby et al. (2000) only proposed an iterative algorithm for discrete low dimensional random variables that obtains a local optimum. The difficulty in generalizing to high-dimensional data is still in the estimation of mutual information. Specifically, when the input variable and encoded representation are both continuous, the mutual information becomes ∞ Amjad & Geiger (2019). For discrete data, if the encoder $q(t|x)$ is injective, there is essentially no loss of information, therefore $I(X;Z)$ is independent of model parameters. Another challenge that prohibited the effectiveness of the approach was discussed in Kolchinsky et al. (2019), where they listed several caveats of IB when X and Y have a deterministic relationship, including the ineffectiveness of the IB Lagrangian and the existence of trivial solutions.

In recent years, many attempts at tackling these problems has been made. To address the issue of estimating mutual information, Alemi et al. (2017) proposes to use a stochastic NN, so that the encoding is surjective. They derive a variational bound on the mutual information and uses Monte-Carlo sampling for optimization. Belghazi et al. (2018) worked with the Donsker-Varadhan representation for estimating mutual information. Yu et al. (2021) incorporates a matrix-based Renyi's entropy functional to parameterize mutual information. Variations to the IB objective has also been explored. Fischer (2020) suggests that the mutual information between the representation and input variable should be replaced with a conditional mutual information, with the target variable as condition since only the *irrelevant* information in the input should be compressed. Strouse & Schwab (2017) showed that by adding another parameter to control the encoder stochasticity, the new objective encourages a deterministic encoder.

In this work, we describe an alternative for optimizing a deterministic encoder. While Shwartz-Ziv & Tishby (2017) only used quantization for estimating mutual information, we propose to actually perform quantization to the latent codes. In this way, different inputs can be mapped to the same quantization bins, yielding finite mutual information. We show in 3.3 that for deterministic encoders, mutual information becomes the entropy of latent codes, which coincides with lossy data compression. Recent advances in neural image compression have introduced powerful entropy estimators for discrete variables. We benefit from this by incorporating an entropy model parameterized by neural networks to estimate the entropy of representations. This continuous relaxation is a differentiable upper bound Shannon entropy, so it can be jointly optimized with the encoder and decoder. The plug-in nature of the entropy model makes it simple to apply to different learning tasks. We evaluated on CIFAR-10 Krizhevsky et al. (2009) for generalization performance of supervised learning. We also apply the entropy constraint to self-supervised contrastive learning tasks. Furthermore, we tested the robustness of our learned models under adversarial attacks.

2 Related Work

Singh et al. (2020) adopts a similar architecture as ours for compressing features with the goal of saving bits, without theoretical discussion of relationship with the IB. Dubois et al. (2021) provided lower bonds on the compression of representations for guaranteeing prediction performance. Lee et al. (2021) applied the conditional entropy bottleneck to contrastive learning tasks. Different from our approach, they used a stochastic encoder and modeled the representations as Gaussians. They followed a variational approach towards optimizing their objective.

3 The Entropy Constrained Information Bottleneck

3.1 Derivation of the ECIB objective

The Information Bottleneck:

$$T* = \arg\max_{T} I(Y;T) - \beta I(X;T) \tag{3}$$

With $\beta \in \mathbb{R}_{\geq 0}$.The Conditional Entropy Bottleneck Fischer (2020) modifies the the second term of the IB objective into a mutual information conditioned on Y, which resulted in more robust perfor-

mance:

$$T^* = \arg\max_T I(Y;T) - \beta I(X;T|Y) \tag{4}$$

To understand the motive, consider the venn diagrams in 1. The red and blue circle represents the entropy of random variable X and Y. The optimal representation z should be the area where z contains all the information in X that is relevant to Y and nothing more, as shown in the graph on the left side. However, if there's no compression term, only maximizing the mutual information between z and Y can result in a representation in the graph on the right side. This satisfies the objective of a maximized $I(Y;z)$ but contains useless information that would effect the generalization and robustness of the model. The compression term in the IB objective minimizes $I(z;X)$, which corresponds to the intersection between the area of X and z in the graph. However, Fischer (2020) pointed out that the actual information that should be minimized is the information that is common between X and z but *irrelevant to* Y, corresponding to the area of the intersection of X and z but with the intersection of z and Y removed. This is just the conditional mutual information between z and X, given Y, resulting in the second term of CEB.

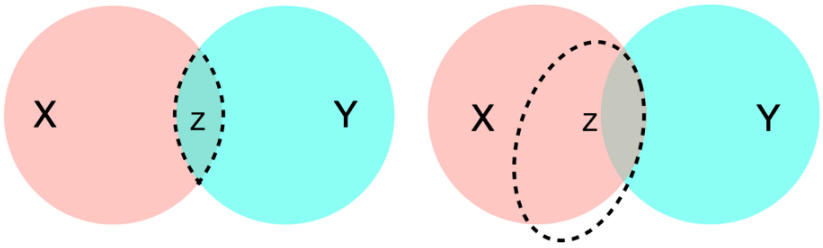

Figure 1: Left: Venn diagram of the optimal representation z obtained by the IB objective. Right: Venn diagram of a possible representation z obtained without the compression term.

By the chain rule of mutual information, the above can be written as:

$$T^* = \arg\max_T I(Y;T) - \beta[I(T;X,Y) - I(T;Y)] \tag{5}$$

For most prediction tasks, the following Markov constraint holds true: $Z \leftarrow X \longleftrightarrow Y$. Given this constraint, all the information about Y in Z is obtained from X. The above objective therefore becomes:

$$\begin{aligned} T^* &= \arg\max_T I(Y;T) - \beta[I(T;X) - I(T;Y)] \\ &= \arg\max_T (\beta+1)I(Y;T) - \beta I(T;X) \\ &= \arg\max_T (\beta+1)(H(Y) - H(Y|T)) - \beta(H(T) - H(T|X)) \end{aligned} \tag{6}$$

Notice $H(Y)$ is a property of the dataset, it can be seen as a constant with respect to T and therefore can be dropped from the function.

$$\begin{aligned} T^* &= \arg\max_T -(\beta+1)H(Y|T) - \beta H(T) + \beta H(T|X) \\ &= \arg\min_T (\beta+1)H(Y|T) + \beta H(T) - \beta H(T|X) \end{aligned} \tag{7}$$

For a deterministic encoder, the conditional entropy (or noise entropy) $H(T|X) = 0$ (Strouse & Schwab (2017)).

$$\begin{aligned} T^* &= \arg\min_T (\beta+1)H(Y|T) + \beta H(T) \\ &= \arg\min_T H(Y|T) + \frac{\beta}{\beta+1} H(T) \end{aligned} \tag{8}$$

Replace $\frac{\beta}{\beta+1}$ with a new parameter $\gamma \in [0, 1)$ gives us the final form of the ECIB objective:

$$T^* = \arg\min_T H(Y|T) + \gamma H(T) \tag{9}$$

Deriving the deterministic version of the original IB would result in almost the same form as the equation above, except in that case the parameter $\gamma = \beta$ and is not confined in $[0, 1)$. For a supervised classification problem, the first term in the EIEB $H((Y|T)$ can be equivalent to optimizing for the average cross-entropy loss. The second term $H(T)$ is just the entropy of the representations. Minimizing this term can be formulated as a source compression problem.

3.2 Entropy Estimation

Recent years, many advances have been made in lossy data compression. This is largely due to the nonlinear approximation ability of neural networks. Traditional codecs such as JPEG and JPEG2000 uses linear transform suchs as DCT and Wavelet transforms to decorrelate signals, perform efficient vector quantization to their transformed coefficients according to hand-designed quantization tables. Ballé et al. (2017) suggested that learning nonlinear transforms as encoder and decoders end-to-end and performing scalar quantization in the transformed space is effectively performing optimal vector quantization in the pixel space, and could yield improved rate-distortion performance. The key ingredient in this framework is a differentiable entropy model that can estimate the marginal distributions of the transformed coefficients (or latent variables), such that lossless entropy coding can be applied. In Ballé et al. (2017), the authors proposed a piece-wise linear function to model the entropy of latent variables, essentially a histogram of probabilities. This serves as a continuous relaxation of the Shannon entropy of the latent variables. In Ballé et al. (2018), a more refined entropy model was introduced to model the dependencies among latent dimensions. Now every latent element is modeled as a Gaussian distribution, and the same factorized model mentioned above is used to model the parameters of the Gaussians. Since the model is learned end-to-end to optimize the rate-distortion trade-off, the entropy model and the encoder function are jointly optimized, encouraging the encoder to produce low-entropy latents. In this work, we adopt this entropy model to estimate the entropy of T and jointly optimize our encoder.

3.3 Overall Implementation

We show the diagram of our model in Figure 2. At first, we pass the input data into out encoder network, and output a representation h. This h will then be used to estimate the Shannon entropy of the quantized version its self. The quantization is necessary because Shannon entropy requires discrete symbols. The quantization step is also the main source of information loss. We follow Ballé et al. (2017) and implement a simple rounding for quantization, relying on the encoder to approximate the optimal transform for scalar quantization. However, this quantization step will be problematic for end-to-end optimization because the gradient of a step function is 0 or infinite everywhere. We use a straight through estimator to enable gradient descent.

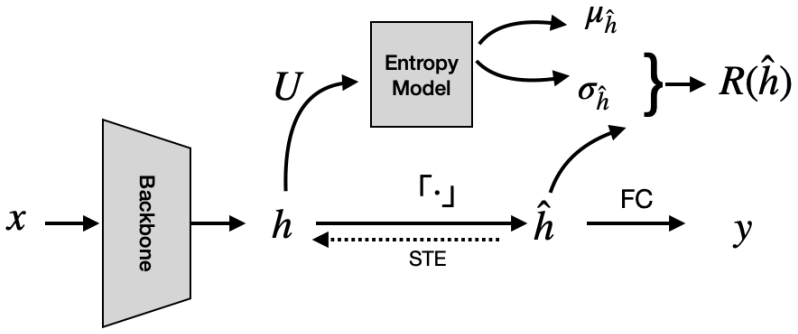

Figure 2: Model Diagram

Quantization error modeling methods such as uniform additive noise. However, in our case the

The quantized representations is then passed trough a fully connected layer to predict the categorical distribution of labels. We used a similar entropy model as in Ballé et al. (2018), except that instead of using convolutional layers, we implemented fully connected layers with ReLU as activation functions. This is because the dataset that we worked on has a smaller scale and spatial information is not of much use in the latent representations. The entropy model estimates the Gaussian parameters of the marginal distribution of each dimension of the quantized latents, and we compute the number of bits needed to code the actual latents using these Gaussians as priors. We use the Lagrangian multiplier in the loss function to control the Rate-Distortion trade-off.

4 EXPERIMENTS

4.1 SUPERVISED CLASSIFICATION

In the supervised classification task, we use the cross-entropy loss a proxy for mutual information between Y and Z since they both represent informativeness. By adding the entropy penalty, we get the loss function for a classifier:

$$L_{sup} = H(Y|g_\theta(f_\phi(X))) + \beta H(f_\phi(X)) \tag{10}$$

Where g_θ is the decoder function that maps a representation to a categorical distribution over labels, and f_ϕ is the encoder function that embeds the input into a representation.

We first trained a supervised learning classification model, and compared the models using different compression rates. We show that with compression rates down to 0.13 bits per dimension, the test accuracy is still comparable and sometimes higher than the model trained without compression.

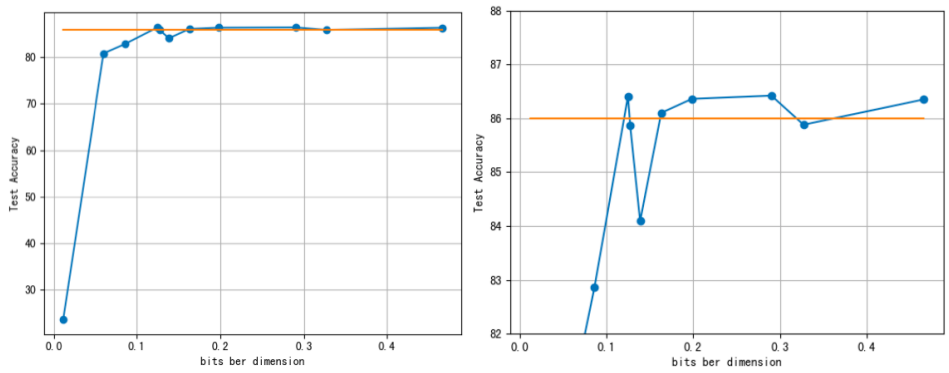

Figure 3: Classification accuracy on the test set

By visualizing the gradient saliency maps Chattopadhay et al. (2018) of selected images, we show that models trained with only cross entropy sometimes overfit to backgrounds like the sky and water, while the compressive model is often able to compress that information out.

4.2 CONTRASTIVE LEARNING

The entropy constraint can also be applied to contrastive learning. It has been shown in Oord et al. (2018) that the contrastive loss based on noise contrastive estimation is a lower bound of the mutual information between randomly augmented views. In this case, the target variable would be another view of the same data. We can apply the deterministic information bottleneck objective to this:

$$\phi* = argmin_\phi - I(X_1|X_2) + \beta H(f_\phi(X_1)) + \beta H(f_\phi(X_2)) \tag{11}$$

Where f_ϕ denotes the encoder function, X_1, X_2 are 2 randomly augmented views of the same image.

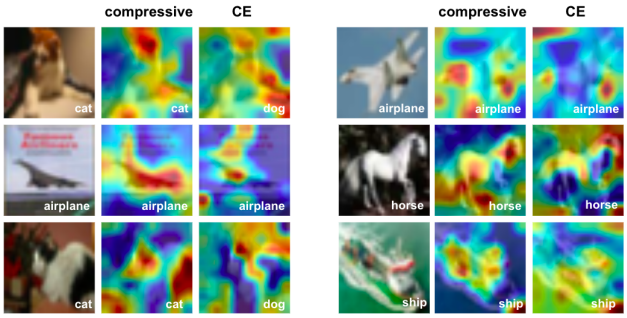

Figure 4: Gradient Saliency Maps

We evaluate the representations on the standard linear classification protocol for contrastive learning. We find that compressive self-supervised representations generalize better than the vanilla Simclr Chen et al. (2020), with a higher test accuracy.

Table 1: Linear Evaluation

Model	**Train Accuracy**	**Test Accuracy**
SimCLR	**71.97**	71.04
Compressive(β=0.05)	.71.51	71.06
Compressive(β=0.1)	71.73	**71.31**
Compressive(β=0.5)	71.14	70.66

4.3 Adversarial Robustness

We evaluate for non-regularized and regularized models. The non-regularized models where trained for 330 epochs, with no data augmentation and weight decay. The regularized models where trained for 100 epochs, took in randomly augmented images at training time, and was applied with a weight decay of 0.0001. CE denotes the model trained with cross entropy loss, different βs correspond to models trained with different compression rates.

Table 2: Adversarial Robustness evaluation of different models.

	Adversarial Attack on non-regularized ResNet-18						
Model	CE	β=0.5	β=1	β=2	β=4	β=8	β=32
Clean	78.25	77.80	78.34	**78.50**	78.41	77.15	74.27
PGD (steps=7)	1.48	**44.29**	40.54	27.28	33.39	36.51	26.68
PGD (steps=20)	1.15	**39.00**	32.34	20.86	19.49	24.95	15.40

	Adversarial Attack on regularized ResNet-18						
Model	CE	β=0.5	β=1	β=2	β=4	β=8	β=32
Clean	85.99	86.36	86.10	**86.41**	85.86	84.09	80.28
PGD (steps=7)	0.02	0.21	0.04	0.17	0.12	0.81	**3.56**
PGD (steps=20)	0.00	0.05	0.00	0.04	0.00	0.54	**1.82**

The reason that we tested for non-regularized models is to show the effect of the entropy constraint on its own. These results show that the entropy constraint is beneficial for more learning robust representations, especially for non-regularized models. The reason why regularized models don't benefit that much from compression might be due to the robustness and generalization trade-off Tsipras et al. (2018), where overfitted models exhibit higher robustness.

4.4 SPARSITY

We continue to investigate properties of the compressive representations and observe strong sparsity. We show in Figure 5 left that with as low as an average of 7 non-zero entries among a 512 dimensional representation vector, the model is still able to achieve performance comparable to non-compressive models.

We can also look at the fully connected layer, which is just a linear transformation matrix that maps the representations to logits. We found that with higher compression, the singular values of the transformation matrix tend to aggregate into fewer larger ones, becoming more approximately low-rank.

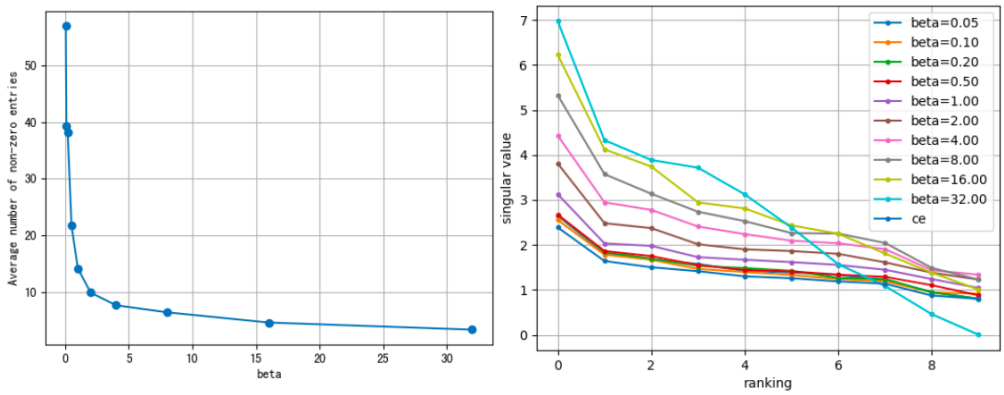

Figure 5: Left: x-axis:beta parameter, corresponding to compression strength; y-axis: the average number of non-zero entries in the learned representations. Right: Singular value distribution of the fully connected layer weight matrix

It's also worth noting that the compressed representation does not necessarily have to be the last layer before the fully connected layer. It could be the output of any intermediate hidden layers. We also applied the entropy penalty on the outputs of the second convolutional layer, and visualize in Figure 6. the sparse feature maps it has learned.

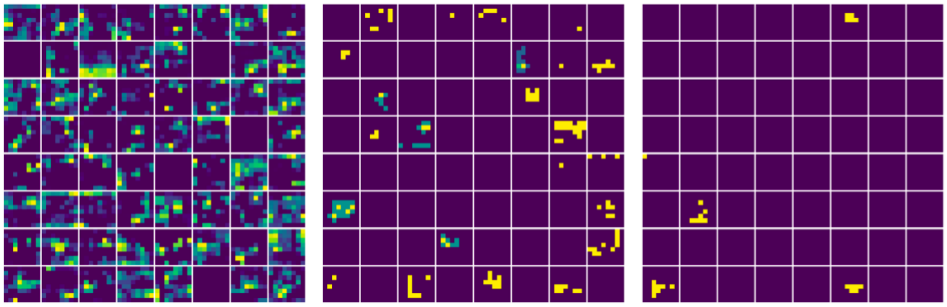

Figure 6: From left to right: the 64 feature maps learned by no-compression, weak compression and strong compression.

5 CONCLUSION

In conclusion, our information theoretical objective along with entropy estimation was able to learn more sparse, robust and generalizable representations compared to no entropy constraint.

For future work we would like to first improve generalization. Even though we have seen comparable or slightly improved test accuracy over non-compressive models, there still remains space for improvement. A direction that would be interesting is to use deep mutual information estimation

Belghazi et al. (2018) for $H(Y|Z)$. Applying out method to a wider range of tasks and data domains is also worth exploring.

REFERENCES

Alexander A. Alemi, Ian S. Fischer, Joshua V. Dillon, and Kevin P. Murphy. Deep variational information bottleneck. *ArXiv*, abs/1612.00410, 2017.

Rana Ali Amjad and Bernhard C Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2225–2239, 2019.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. *ArXiv*, abs/1611.01704, 2017.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Yann Dubois, Benjamin Bloem-Reddy, and Chris J. Maddison. Lossy compression for lossless prediction. In *NeurIPS*, 2021.

Ian S. Fischer. The conditional entropy bottleneck. *Entropy*, 22, 2020.

Artemy Kolchinsky, Brendan D. Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios. *arXiv: Machine Learning*, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34:19538–19552, 2021.

Stephan Sloth Lorenzen, C. Igel, and M. Nielsen. Information bottleneck: Exact analysis of (quantized) neural networks. *ArXiv*, abs/2106.12912, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu S. Advani, Artemy Kolchinsky, Brendan D. Tracey, and David D. Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019, 2018.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *ArXiv*, abs/1703.00810, 2017.

Saurabh Singh, Sami Abu-El-Haija, Nick Johnston, Johannes Ball'e, Abhinav Shrivastava, and George Toderici. End-to-end learning of compressible features. *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3349–3353, 2020.

DJ Strouse and David J. Schwab. The deterministic information bottleneck. *Neural Computation*, 29:1611–1630, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *ArXiv*, physics/0004057, 2000.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Xi Yu, Shujian Yu, and José Carlos Príncipe. Deep deterministic information bottleneck with matrix-based entropy functional. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3160–3164, 2021.