Entropy Constrained Compressive Representation Learning

1st Lingyu Zhang
Department of Electrical Engineering
Columbia University
New York, USA
{lz2814}@columbia.edu

***Abstract*—The information bottleneck method [16] is an information theoretical objective for generalizable and robust representation learning. However, it has been critiqued to be infeasible for high-dimensional data due to the difficulty in estimating mutual information. In this work we propose an deterministic information bottleneck with neural entropy estimation. We show that for a deterministic encoder, the mutual information estimation problem can be formulated as a source compression problem, where we then integrate variational entropy estimation commonly used in neural image compression to jointly optimize the encoder and entropy model. We evaluate our method on supervised and self-supervised learning tasks for generalization, and on against adversaraial attack for robustness. We show that out 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/Mikeyboiii/cSCL**

***Index Terms*—Representation Learning, Information Bottleneck, Entropy Estimation**

I. Introduction

The information bottleneck (IB) [16] introduced an objective for representation learning that trades off its informativeness of the target variable and its compressiveness with respect to the input variable, enforcing it to be the minimal sufficient statistics of the input:

$$\max_{Z} \text{IB}(X, Y, Z) = I(Y; Z) - \beta I(Z; X)$$

Where the task is to predict variable Y from X, Z is a set of features selected from X, $I(\cdot;\cdot)$ is the mutual information between two random variables. The first term encourages Z to contain sufficient information for predicting Y, while the second term forces Z to be forgetful of X. By optimizing this Lagrangian, Z is learned to contain only the information from X that is predictive of Y, resulting in good generalization and robustness. However, the authors only proposed an iterative algorithm for discrete random variables that obtains a local optimum. To challenge in generalizing to high-dimensional continuous distributions lies in the estimation of mutual information, which is generally difficult.

In recent years, many attempts at approximating or circumventing this problem has been made. Variational inference methods has been advanced as a result of neural networks' capability to approximate complex distributions. [1] parameterizes distributions with neural networks and uses Monte-Carlo sampling to optimize a lower bound on the IB objective. [7] uses variational inference to learn disentangled representations in a self-supervised fashion. Variations to the IB has also been extensively explored. [5] suggests that the mutual information between the representation and input variable should be replaced with a conditional mutual information, since only the irrelevant information in the input should be minimized. [18] incorporates a matrix-based Renyi's entropy functional to parameterize mutual information. [15] showed that by adding another parameter to control the encoder stochasticity, the new objective encourages a deterministic encoder.

In this work, we propose an alternative for optimizing a deterministic encoder. Inspired by recent advances in learned image compression, we incorporate 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 MNIST [9] and CIFAR-10 [8] 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.

II. Methods

A. The Deterministic Conditional Entropy Bottleneck

In this section, we derive our deterministic information bottleneck objective from the the conditional entropy bottleneck and the deterministic entropy bottleneck.

The Conditional Entropy Bottleneck (CEB) introduced in [5] modifies the the second term of the IB objective into a mutual information conditioned on Y, which resulted in more robust performance:

$$\max_{Z} \text{CEB}(X, Y, Z) = I(Y; Z) - \beta I(Z; X, Y)$$

To understand the motive, consider the venn diagrams in Figure 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 satisfy 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, Fishcer et al. 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.

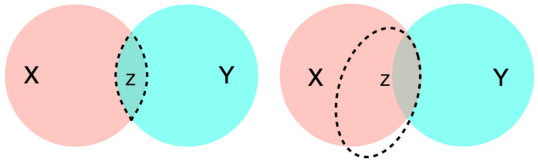

Fig. 1. Left: Venn diagram of the optimal representation z obtained by the IB objective. Right: Venn diagram of a possible representation z obtained without the compression term.

We then consider the deterministic information bottleneck [15]. We can first write mutual information between Z and X as the entropy of Z subtracted by the conditional entropy of Z given X.

$$I(Z;X) = H(Z) - H(Z|X)$$

This is almost a definition of mutual information, which is given X, how much uncertainty can we deduct from Z. This conditional entropy term is called noise entropy in communication theory, which measures the stochasticity in the mapping from X to Z. In the extreme case, Z can be a deterministic function of X. Then this term would be 0, suggesting that given the value of X, we know the value of Z with probability of 1. But the authors pointed out that naively replacing this term in the information bottleneck objective will result in an objective that cannot be solved by variational calculus because if you the encoder does not appear explicitly in the derivative of the new objective. What they did instead is introduce another Lagrangian to control the noise entropy.

We now show that for deterministic encoders, the CEB and the IB is equivalent. We can first rewrite the second mutual information term in the CEB in a similar way:

$$I(Z;X,Y) = H(Z) - H(Z|X,Y)$$

Since $X \rightarrow Z$ is deterministic, $H(Z|X,Y)$ is again 0. Pluging this back into the CEB objective, we get:

$$\max_Z \text{CEB}(X,Y,Z) = I(Y;Z) - \beta[H(Z) - H(Z|X,Y)] \tag{1}$$

$$= H(Y) - H(Y|Z) - \beta H(Z) \tag{2}$$

$$= -H(Y|Z) - \beta H(Z) \tag{3}$$

We can drop $H(Y)$ in (2) because it is constant with respect to Z. Note that this is exactly the same as IB with a deterministic function because conditioning on Y no longer matters when $X \rightarrow Z$ is deterministic. Since Z is now just a function of X, we can write the optimization in terms of the function's parameters:

$$\phi* = argmin_\phi H(Y|f_\phi(X)) + \beta H(f_\phi(X))$$

Now the challenge lies in estimating the entropy of the latent variables $Z = f_\phi(X)$ and optimizing it with respect to the encoder function. We propose that this can be learned end-to-end with neural networks.

B. 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. [2] suggested that learning a nonlinear transform end-to-end and performing scalar quantization in the transform space could yield improved rate-distortion results. The key ingredient in this framework is a differentiable entropy model that can estimate the distributions of the transformed coefficients (or in latent variables), so that lossless entropy coding can be applied. In [2], 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, and upper bounds it since non-exact approximation of symbol distributions will result in sub-optimal code. In [19], 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 Z and jointly optimize our encoder.

C. Supervised Learning

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:

$$L_{sup} = H(Y|g_\theta(f_\phi(X))) + \beta H(f_\phi(X)) \quad (4)$$

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.

D. Contrastive Learning

The entropy penalty can also be applied to contrastive learning. It has been shown in [17] 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:

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

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

III. Experimental Setup

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 [2] 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. 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 [19], 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.

To evaluate out method, we first conducted experiments with supervised classification on CIFAR-10 [8]. We use a ResNet-18 [6] as backbone for the encoder network, and use a mean-scale hyperprior [19] for entropy modeling. We used a batch size of 128, an initial learning rate of 5e-3 with cosine annealing decay strategy. A weight decay of 1e-4 was applied. We trained for a total of 100 epochs.

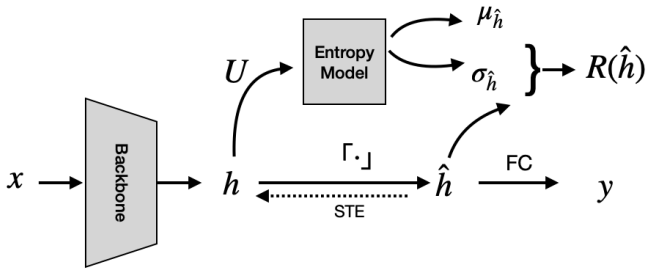

Fig. 2. Model Diagram

For the representations learned with a contrastive loss, we evaluate them on the linear evaluation protocol of image classification task and compare with [4]. We used a temporature of 0.07 for the contrastive loss, used a batch size of 256 for training. We trained with a learning rate of 3e-4 for 200 epochs, and another 200 epochs with 3e-5. The representations where projected to a 128 dimensional vector for the contrastive loss.

We evaluate robustness under adversarial attack. Following [11], we applied projected gradient descent attacks with an ϵ budget of 8/255, and step size of 2/255 for 7 and 20 steps.

IV. Results

A. Supervised Learning Accuracy

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.

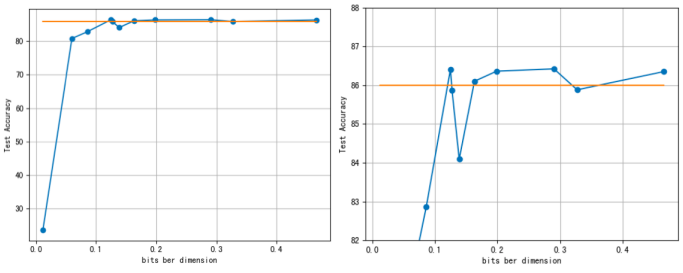

Fig. 3. Classification accuracy on the test set

By visualizing the gradient saliency maps [13] 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.

B. Linear Evaluation of Contrastive Representations

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 [4], with a higher test accuracy.

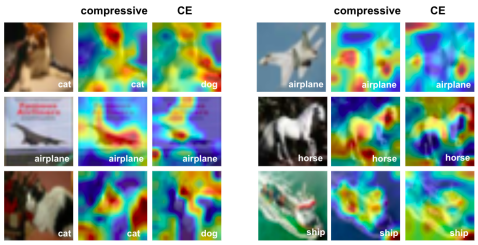

Fig. 4. Gradient Saliency Maps

TABLE I
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

C. 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 II
ADVERSARIAL ROBUSTNESS EVALUATION OF DIFFERENT MODELS.

Model	Adversarial Attack on non-regularized ResNet-18						
	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

Model	Adversarial Attack on regularized ResNet-18						
	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 [12], where overfitted models exhibit higher robustness.

D. 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.

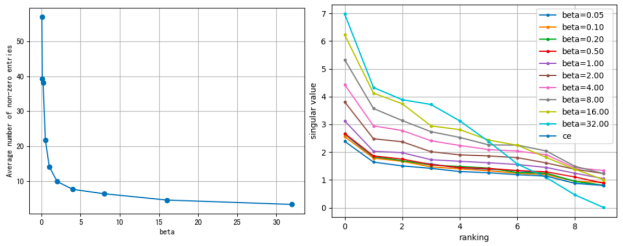

Fig. 5. Left: x-axis:beta parameter, corresponding to compression strength; y-axis: the average number of non-zero entries in the learned representations. Right: Singular value distribution of the fully connected layer weight matrix

It's also worth noting that the compressed representation doesn't 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.

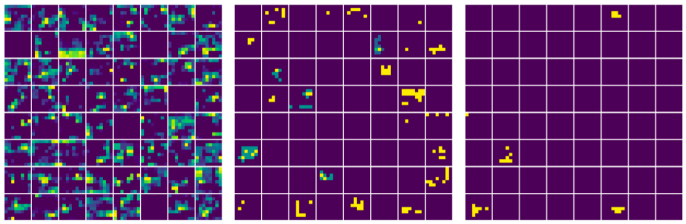

Fig. 6. From left to right: the 64 feature maps learned by no-compression, weak compression and strong compression.

V. RELATED WORKS

[10] 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. Our framework is similar to [14], where they also explored the joint task of compression and classification. However, they aimed at using compression for space efficient features, and did not make theoretical analysis in terms of connection to the information bottleneck.

VI. 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 [3] for $H(Y|Z)$. Applying out method to a wider range of tasks and data domains is also worth exploring.

References

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

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

[3] Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron C. Courville. Mine: Mutual information neural estimation. ArXiv, abs/1801.04062, 2018.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. ArXiv, abs/2002.05709, 2020.

[5] Ian S. Fischer. The conditional entropy bottleneck. Entropy, 22, 2020.

[6] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[7] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In ICLR, 2017.

[8] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/ kriz/cifar.html.

[9] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

[10] Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John F. Canny, and Ian S. Fischer. Compressive visual representations. In NeurIPS, 2021.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. ArXiv, abs/1706.06083, 2018.

[12] O.and Bueno G. Pedraza, A.and Deniz. On the relationship between generalization and robustness to adversarial examples. Symmetry, pp. 13, 81, 2021.

[13] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128:336–359, 2019.

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

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

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

[17] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. ArXiv, abs/1807.03748, 2018.

[18] Xi Yu and 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.

[19] Ballë, J., Minnen, D.C., Singh, S., Hwang, S.J., Johnston, N. (2018). Variational image compression with a scale hyperprior. ArXiv, abs/1802.01436.

VII. Appendix

A. Discussions of the suboptimal generalization performance

We find that models regularized with an entropy penalty hasn't been able to largely improve generalization. Possible causes include the non-equivalence of the conditional entropy $H(Y|Z)$ and the cross entropy loss:

$$L_{Cond} = H(Y|Z) = -\sum_z p(z) \sum_y p(y|z) \log p(y|z) \quad (5)$$

$$= -\sum_z p(z) \log p(y|z) \quad (6)$$

$$L_{Cross} = \sum_z H(Y, \hat{Y}(z)) = -\sum_z \log p(y|z) \quad (7)$$

Where the conditional entropy is marginalized over the distribution of representations, and the cross-entropy is only an average. For high-dimensional representations, all distinctive zs only appear once, so there exists a gap between the two objectives.

For contrastive learning, as it is shown by [17], the contrastive loss is also not a tight bound on the mutual information between views:

$$I(z_1, z_2) \geq \log(N) - L_{cont}(z_1, z_2) \quad (8)$$

Where N is the number of contrastive samples in a batch. Due to time and computation limitations, we used $N = 256$ which is a relatively small batch. It has been shown in [4] that the batch size directly effects the quality of representations. If the representations are not informative enough on its own, it's unlikely that adding a compression term would improve performance.

Another possible reason is the difficulty in optimizing a joint loss. The compressive network is more complex than a single classification model so it is likely that it requires longer training time and different training strategies. We've found that emprically, the entropy model converges slower than the classifier. In the early stages of experiments, we trained models for far less epochs and observed decreasing performance when the entropy penalty is applied. By visualizing the loss curve, it was clear that different models converged at different speed so training for the same amount of time steps is limiting the power of compressive models. This could also explain the nonmonotonic relationship between β and accuracy. However, for fair comparison, we did not try different strategies for different models.

Furthermore, there is also the possibility of the occurence of overfitting in the entropy model its self, where the model memorizes marginal distributions of the training set and cannot generalize well to the test set.