On the Impact of Perceptual Compression on Deep Learning

Gerald Friedland, Ruoxi Jia
University of California, Berkeley
fractor@eecs.berkeley.edu, ruoxijia@berkeley.edu

Jingkang Wang
University of Toronto
wangjksjtu@gmail.com

Bo Li
University of Illinios, Urbana Champaign
lxbosky@gmail.com

Nathan Mundhenk
Lawrence Livermore National Lab
mundhenk1@llnl.gov

Abstract

This paper proposes a fundamental answer to a frequently asked question in multimedia evaluation and data set creation: Do artifacts from perceptual compression contribute to error in the machine learning process and if so, how much? Our approach to the problem is an information reinterpretation of the Helmholtz free energy formula to explain the relationship between content and noise when using sensors (such as cameras or microphones) to capture multimedia data. The reinterpretation guides a bit-measurement of the noise contained in images, audio, and video by combining a classifier with perceptual compression, such as JPEG or MP3. Our experiments on CIFAR-10, ImageNet, and CSAIL Places as well as Fraunhofer's IDMT-SMT-Audio-Effects dataset indicate that, at the right quality level, perceptual compression is actually not harmful but contributes to a significant reduction of complexity of the machine learning process. That is, our noise quantification method can be used to speed up the training of deep learning classifiers significantly while maintaining, or sometimes even improving, overall classification accuracy.

1 Introduction

As datasets for multimedia [17] grow larger and larger and become more difficult to handle, performing machine learning on perceptually compressed data has become increasingly mainstream [11]. In the past practice of feature extraction, however, many signal processing communities had established a firm rule that features should be extracted on uncompressed input only. Today, deep learning systems have replaced many feature-based systems for computer vision and audition tasks. Still, the preference is often to develop deep learning systems on high-quality image and audio – even if it might result in a more complex machine learning task. This is, due to the lack of a clear understanding of the relationship between lossy compression and deep learning, researchers understandably choose to "play it safe". In this article, we hope to contribute to this understanding.

In fact, the interaction of perceptual compression and machine learning is not at all very well studied. Dodge et al. [4] demonstrate that the performance of deep neural networks is "surprisingly" robust to artifacts introduced by perceptual compression as apposed to other types of image distortions such as blurring and random noise. As a note, perceptual compression has also proved useful for improving models' robustness against adversarial example-signals that are intentionally made close to natural multimedia signals but misclassified by models [3].

In this article, we present a physical model that describes the interaction of sensor data with machine learning along with empirical evidence to verify the theoretical hypotheses. We show how, using the methodology employed in this paper, it is possible to estimate the amount of noise versus the amount of content in sensor data by deriving the expected shape of the measurement results in Figure 2.

2 Physical Model

In thermodynamics, the Helmholtz free energy is a thermodynamic potential that measures the "useful" work obtainable from a closed thermodynamic system at a constant temperature and volume. The Helmholtz free energy is defined as

$$A \equiv U - TS, \tag{1}$$

where A is the Helmholtz free energy, U is the internal energy of the system, T is the absolute temperature of the surroundings in Kelvin, and S is the Boltzmann entropy of the system. The Boltzmann entropy is given as $S = -k \log P$ where k is a constant and P is the probability of observing a micro state. This assumes each state is equiprobable. While unproven, since it's original publication in 1882 [18], it has been found many times that this formula is generally useful

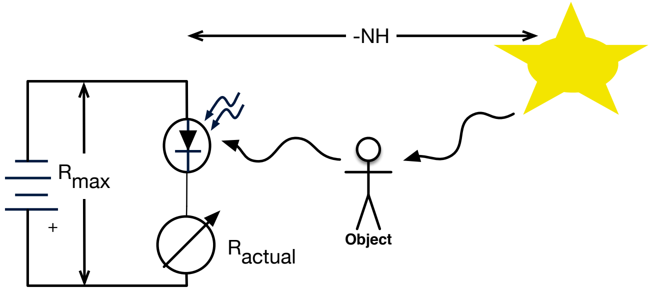


Figure 1: Our interpretation of the Helmholtz free energy equation $A \equiv U - TS$ applied to camera sensors. Potential energy of the battery (U) is converted into kinetic energy A (electron flow) in response to light hitting the photo sensor. The free energy A is therefore the maximum electron flow observable in the circuit controlled by the photo sensor in response to the energy loss due to the collision of a sunbeam with a reflecting object. Assuming a constant sampling rate, the measured signal strength R_{actual} corresponds to A divided by the sampling period length. R_{max} corresponds to U in the same way. We reinterpret the energy loss TS in the same time period as NH (read: noise times information content).

to describe the influence of an externality on a thermodynamically closed system (a system that exchanges energy but not matter with its surroundings). Most importantly, the Helmholtz formula is often used to define fundamental equations of physical state from which key variables are interpreted as probability distributions. Inspired by this notion, we reinterpret the formula for the purpose of sensor measurements as they frequently occur in the field of Multimedia Computing.

We apply the formula as illustrated in Figure 1. Potential energy of the battery (U) is converted into kinetic energy A (electron flow) in response to light hitting the photo sensor. The solar energy, however, is reduced by hitting objects it reflects from, changing both frequency (chrominance) and amplitude (luminance) of the light wave. While modern cameras have 3 channels, this is three different photo sensors, to distinguish the different energy losses conforming to human vision, for simplicity we only assume one channel here. The free energy A is therefore the maximum electron flow observable in the circuit controlled by the photo sensor. A pixel is usually treated as a signal which is, physically speaking, a power while energy is power multiplied by time. Treating measurements as signals is convenient for computation but in reality, any sensor requires activation energy, which is, a signal through time. To resolve this impreciseness, we assume that the measurement interval (often referred to as frame rate, capture time, or sampling rate) is constant. This is usually true for multimedia capturing devices such as photo and video cameras, or digital microphones. The assumption is convenient as it makes the signal treatable as an energy divided by a constant. We assume this constant to be 1. We denote the measured pixel as R_{actual} and it therefore corresponds to A divided by 1. The reference white R_{max} corresponds to U in the same way and is the maximum value that can be measured. If the photo element was directly pointed at the sun, then $R_{actual} = R_{max}$. We denote the energy loss TS as NH. This energy loss is the result of the photo sensor being influenced by an externality: As the sunlight is reflected from the surface of a certain chemical composition, it loses energy. According to Helmholtz, this loss is the comprised of two factors: 1) An entropy term S (unknown a-priori to the photo sensor circuit), which we call H and 2) temperature T, which we generalize as the noise constant N. N is an unknown scalar that characterizes the noise effects in the measurement process (for example, for image sensors this includes everything from camera lens aberration to dead pixels but also actual thermodynamic variation), and H is the information captured by the sensing process. We replace S, which assumes equi-distribution of states, with H, the Hartley entropy [13].

This is, we re-purpose the Helmholtz formula as follows:

$$R_{actual} = R_{max} - NH, \tag{2}$$

where R_{actual} is the information content of the actual sensor reading (for example, 4 bits), R_{max} is the information content of the maximum sensor reading (for example 8 bits), NH is the signal loss characterized by the expected minimum description length of that loss H times a noise scalar N. Note that since we don't use Boltzmann's constant k anymore, we cannot measure N in Kelvin.

In the case of audio, R_{max} captures the information content of the maximum signal strength that the microphone can record to generate one sample. H is the information contained of the sample in bits and N is noise. Similar analogies can be made for other sensors.

3 Perceptual Compression

Most media distributed for typical multimedia analysis benchmark sets are compressible using JPEG or MPEG. The degree of compression can usually be adjusted, allowing one to control the trade off between storage size and media quality. Compression fits our physical model as follows: The information content of a pixel value R_{actual} in each Y, U, or V dimension is represented as described in Equation 2. Since R_{max} is constant (usually 8 bit), we disregard R_{max} as well as the sign before the term NH and observe that the a pixel is composed of information content measured as H multiplied by a noise factor N.

We can therefore obtain an estimate of the noise factor via the following equation:

$$\frac{NH}{N_{approx}} = H_{approx}, \tag{3}$$

where N_{approx} is the approximated noise factor. H_{approx} is then the approximated information content of the pixel and can therefore be compressed using lossless compression. The working hypothesis for our article, and as indicated by related work, is that data quantized to information content H_{approx} can be modelled better using any encoder, including a neural network (we assume the neural network as encoder model suggested by [9]).

3.1 Measuring N_{approx}

In order to get an estimate of the content portion relevant for classification, we train a classifier with identical input images, except different quality levels of JPEG. We then expect the typical training error of the classification to drop steeply at the point where the quantization gets too high. That is, at the point where the chosen JPEG quality level q implies $N_{approx} > N$. Choosing an optimal q, however, should lead to overall less training time because the number of parameters for the machine learner can be reduced as most of the complexity induced by noise does not have to be modeled. Since we do not know the underlying distribution that underlies the measures H and H_{approx} and we also do not know how the distribution that the encoded weights have at a given stopping point of training, we are forced to assume the worst case, this is H and H_{approx} being uniformly distributed. This is, we reduce the complexity of the sensor signal uniformly, erasing both noise and content at the same time. This makes the expected accuracy proportional to our estimation of the information content $NH = -\log_2 P$, where P is the probability of a concrete sensor reading (e.g., $P = \frac{1}{2^8}$ for 8 bit sensors). As a consequence, the probability of modeling a sensor reading correctly for a large amount of samples (law of large numbers) equals the average accuracy. This allows us to draw the curve in Figure 2. It is generated as $Accuracy = c * \log(q)$ where c is a scaling constant and q is quality in percent. $\log_2(q)$ is of course proportional to the upper bound on the information content of a compressed image (see also: Fano's inequality [2]). The point where the curve increases drastically would be N_{approx}. If $N_{approx} > N$, the quantization will, intuitively speaking, “cut into the content” and destroy information contained in H. Note that this upper-limit argument is still valid for noise from perceptual compression that is not independent random noise or images whose pixel values are not uniformly distributed. Dependent noise and patterns in the image will only give us better results as machine learning can exploit the patterns created by the dependency on noise or by recognizing patterns. In other words, the curve should become less smooth and the point N_{approx} more clearly visible.

Acoustic perceptual compression works conceptually similar to visual compression [6]. However, the ear is in general more sensitive to noise distortion than the eye. Therefore, a more accurate version of the DCT, the mDCT [8], is usually applied and the quantization matrices are not linear but tuned to the human auditory system using empirical measurements. That is, our theory remains unchanged for acoustic signals but we expect our hypothesized curve to be less accurate due to the non-linear quantization (compare also Figure 5a).

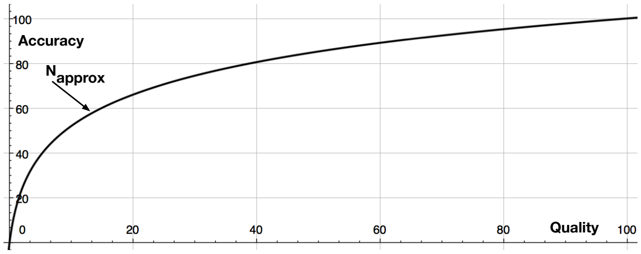


Figure 2: Hypothesized accuracy/quality curve as per the model derived from the Helmholtz Free Energy equation 2. The Y-axis shows a hypothesized accuracy and the X-axis shows a decreasing quantization factor. The sharp drop should occur at the point where the quantization approximates the noise most exactly. Note: The theory only predicts the upper-limit slope and the axis scaling is arbitrarily chosen.

4 Experimental Results

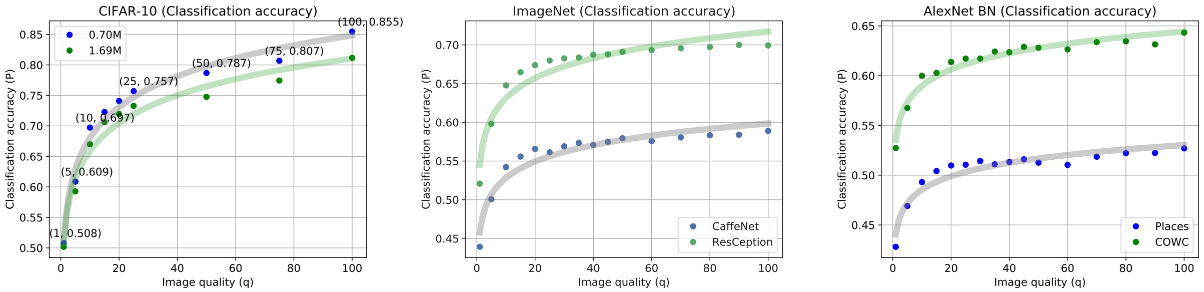
Our experiments had two goals. First, we want to confirm the shape of the curve predicted in Figure 2 and, as a result, being able to confirm the possibility of measuring N_{approx}. Second, we wanted to see if the knowledge of N_{approx} contributes to a reduction of complexity of machine learning models. Note that, due to space constraints, this article only presents a subset of all our experiments. Therefore, as outlined in Section 5, the full set of experiments is available online for reproducibility.

4.1 Image Classification Accuracy

For our image experiments, we separately reproduced the theoretic results on four different image datasets, namely CIFAR-10 [7], ImageNet [12], Places [19] and COWC [10]. We apply different levels of JPEG compression to correlate the compression ratio with the classification accuracy for each of the respective tasks. The results are demonstrated in Figure 3.

4.2 Image Classification Complexity

Our experiments were designed to reduce the complexity of networks while maintaining high performance. For CIFAR-10, we explored six model architectures (A to F) with different number of parameters (0.7M to 1.69M), covering several classic architectures of deep neural networks.



(a)

(b)

(c)

Figure 3: Classification accuracy as a function of the JPEG compression quality q on (a) CIFAR-10 averaged using 6 different classification architectures, b) on ImageNet using two CaffeNet and ResCeption, and (c) on Places and COWC using AlexNet. The dots indicate actual results and the line is the curve as theorized.

The models are able to handle data of different complexity based on the number of parameters used. They are:

- A: Architecture similar to All Convolutional Net [15], where no fully connected layers are employed, but replace first three convolutional layers with VGG [14] setting (channels 32, 64, 128). Size of parameters: 0.70M. (Blue dots in Figure 3a)
- B: Multiple fully connected layers following convolutional layers are added based on A. Size of parameters: 1.08M
- C: Extra convolutional group (three 128 channel convolutional layers) is extended based on A. Size of parameters: 1.14M
- D: Both convolutional group and multiple fully connected layers are added based on A. Size of parameters 1.28M
- E: Multiple fully connected layers with larger units are adopted compared to B. Size of parameters: 1.62M
- F: More fully connected layers are extended compared to E. Size of parameters: 1.69M (Green dots in Figure 3a)

Figure 4 visualizes different effects of parameter reduction methods, and based on the comparison, we used A to demonstrate the compression effects under different ratios. A similar trend can be observed for different parameter reduction methods.

All image results, small scale and large scale, are consistent with Figure 2 and suggest that a significant cut down on bits only has marginal impact on accuracy until a certain threshold is reached. Since our approach is not image specific and therefore suggests cross-modal validity, we also performed experiments on audio data. These are outlined as follows.

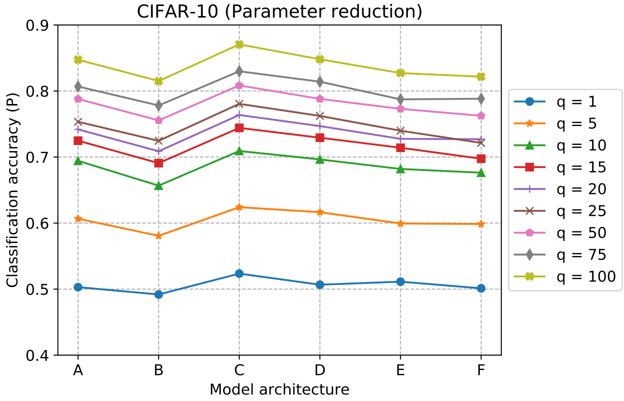


Figure 4: Left: Classification accuracy on CIFAR-10 as a function of number of parameters for different levels of JPEG compression (q value).

4.3 Audio Classification

For the audio experiments, we leveraged the Fraunhofer IDMT-SMT-Audio-Effects dataset [16], which is a large database for automatic detection of audio effects in recordings of electric guitar and bass and related signal processing [1]. This dataset contains 55044 uncompressed WAV files (44.1 kHz, 16 bit, mono) with single recorded notes. In our experiments, we explored a subset of 12 classes of different audio effects and 20592 monophonic guitar samples in total, with 75% for training and 25% for testing.

We perform compression on the input WAV files using the open-source MPEG-Audio Layer 3 implementation LAME [2]. In contrast to JPEG, LAME is not parameterized using a quality level q but using a target bitrate. We then generated mel-spectrograms of all audio files and the number of mel coefficients used in spectrograms is set to 96.

[1] www.idmt.fraunhofer.de/en/business_units/m2d/smt/audio_effects.html

[2] lame.sourceforge.net

Similar to the image experiments, we explored six model architectures (A to F) with different numbers of parameters (0.10M to 3.72M) for the audio classification task. Our performance is comparable to the baseline reported in [16] where a Support Vector Machine is trained. We defined two loops: 1) convolutional layer loop (Conv-loop) that is composed of a convolutional layer, an ELU [1] nonlinear activation layer, a max-pooling layer and a dropout layer; 2) fully connected layer loop (FC-loop) that is composed of a fully connected layer with 128 units and a dropout regularization layer. For better performance, the dropout rates were set to 0.5 and 0.6 in two loops, respectively.

- A: Conv-loop $\times 3$ and no FC-loop. Size of parameters: 0.10M (Blue dots in Figure 5a)
- B: Conv-loop $\times 4$ and FC-loop (128). Size of parameters: 0.17M
- C: Conv-loop $\times 3$ and FC-loop (64-128). Size of parameters: 0.43M (Green dots in Figure 5a)
- D: Conv-loop $\times 3$ and FC-loop (128). Size of parameters: 0.81M
- E: Conv-loop $\times 3$ and FC-loop (128-128). Size of parameters: 0.82M
- F: Conv-loop $\times 2$ and FC-loop (128). Size of parameters: 3.72M

5 Conclusion and Future Work

As can be observed from Figures 3a and 5a, both image and audio experiments follow the trend predicted in Figure 2. That is:

1. The classification accuracy on real data under different compression ratios exactly follows the trend of the hypothesized accuracy curve derived from Equation 2. As a consequence, the empirically measured N_{approx} matches the theoretically calculated N_{approx} well for both image and audio datasets.
2. As predicted, if the quantization level is smaller than N_{approx}, perceptual compression does not seem to affect the classification accuracy significantly. For images, the sweet spot seems to be at $q = 20$ which is the equivalent of 1.4 bits per pixel. Audio shows similar behavior, however, it seems harder to find a concrete sweet spot.
3. Due to the reduction of complexity of the input, neural networks are able to achieve similar classification accuracy using fewer parameters. A smaller number of parameters implies a higher probability to have training converge to a better accuracy faster. As a consequence, our models with a small number of parameters (blue dots) can even achieve a higher classification accuracy than those with a larger number of parameters (green dots).

In conclusion, our empirical results indicate that a quantification and subsequent quantization of the noise content as outlined in this article is useful to reduce the complexity of machine learning: By controlling the level of perceptual compression, we are able to both achieve high learning utility and reduce training complexity. On the other hand, passing pixels unfiltered into a deep learning mechanism therefore means that, before the machine learner can recognize patterns of pixels, one needs to reduce most of the noise before one can get to the significant information per pixel.

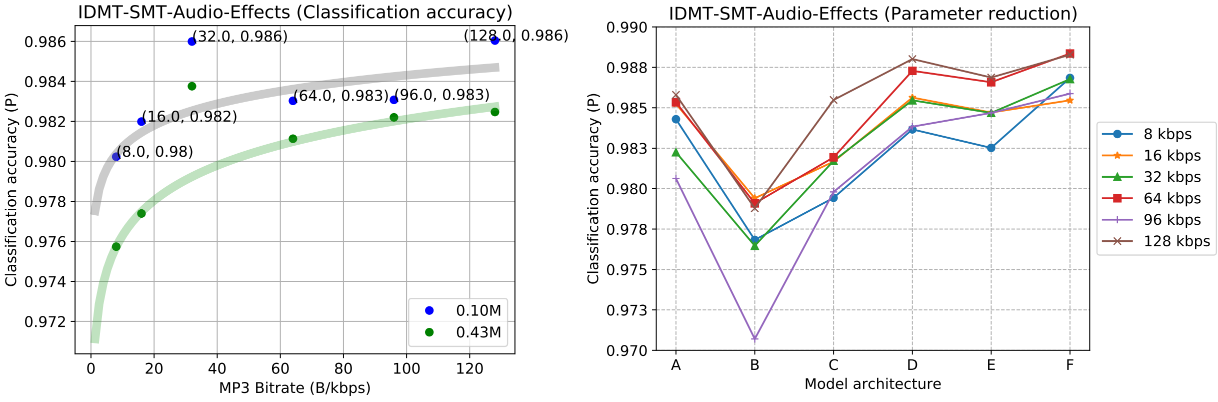
In general, it is easy to take pixels or audio samples for granted and not ask where they come from. However, there is a chain of production that creates the content we are using to investigate machine learning algorithms on multimedia data. In order to understand that chain we sometimes need to go back to the physical fundamentals. Here, applying what is known from physics to our standard methods in machine learning and multimedia computing allowed us to measure the signal to noise ratio of the sensor reading which we used to optimize our machine learning process.

Future work in this fundamental area could use measurement tools like ours to explore cross media boundaries. For example, to quantify the average number of bits needed to distinguish a dog from a cat in images vs in audio vs text. We also speculate that measuring the noise content of images can help explain and identify adversarial examples such as described in [5]. Overall, we hope that our paper contributes to the fundamentals of our field and encourages other researchers to favor measurements over tuning hyper parameters.

All our experiments are available for reproduction at: https://github.com/wangjksjtu/Helmholtz-DL

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. It was partially supported by a Lawrence Livermore Laboratory Directed Research & Development grant (18-ERD-021, “Explainable AI”). IM release number LLNL-CONF-751806. This work was also supported in part by the Republic of Singapore’s National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore. Furthermore, this work has also supported by the National Science Foundation under award numbers CNS-1238959, CNS-1238962,



(a)

(b)

Figure 5: Classification accuracy (a) on the audio classification task as a function of the MP3 compression ratio (relative bitrate). The results of classifier setup A and C are blue and green dots, respectively. The shadow curves represent the properly scaled version of the theoretical curve from Figure 2. (b) shows the classification accuracy at different bitrates under different architectures (A to F).

CNS-1239054, CNS-1239166, CNS 1514509, and CI-P 1629990. Any findings and conclusions are those of the authors, and do not necessarily reflect the views of the funders. We want to cordially thank Alfredo Metere, Jerome Feldman, Kannan Ramchandran, and Bhiksha Raj for their insightful advise.

References

[1] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289, 2015.

[2] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[3] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

[4] S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, pages 1–6. IEEE, 2016.

[5] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 1, 2017.

[6] G. Friedland and R. Jain. *Multimedia Computing*. Cambridge University Press, 2014.

[7] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[8] D. Le Gall. Mpeg: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–58, 1991.

[9] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003.

[10] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. *CoRR*, abs/1609.04453, 2016.

[11] K. Ni, R. Pearce, K. Boakye, B. Van Essen, D. Borth, B. Chen, and E. Wang. Large-scale deep learning on the yfcc100m dataset. *arXiv preprint arXiv:1502.03409*, 2015.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

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

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[15] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

[16] M. Stein, J. Abesser, C. Dittmar, and G. Schuller. Automatic detection of audio effects in guitar and bass recordings. In *Audio Engineering Society Convention 128*, May 2010.

[17] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[18] H. von Helmholtz, J. W. Hittorf, and J. D. Waals. *Physical Memoirs Selected and Translated from Foreign Sources*. Taylor & Francis, 1888.

[19] B. Zhou, A. Khosla, À. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016.