FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Martin Grätzer

Neural Networks and Knowledge Distillation

Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis:	Mgr. Ondřej Týbl, Ph.D.
Study programme:	Financial and Insurance Mathematics

Prague 2025

I declare that I carried out this master thesis on my own, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

Dedication.

Title: Neural Networks and Knowledge Distillation

Author: Bc. Martin Grätzer

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Ondřej Týbl, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Abstract.

Keywords: neural networks machine learning knowledge distillation KL divergence

Název práce: Neuronové sítě a destilace znalostí

Autor: Bc. Martin Grätzer

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Ondřej Týbl, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Abstrakt práce přeložte také do češtiny.

Klíčová slova: neuronové sítě, strojové učení, destilace znalostí, KL divergence

Contents

Introduction

In the recent years we have experienced a remarkable surge in artificial intelligence (AI). This rise has been fueled by an increase in computational power, making the creation of more powerful and complex models feasible. However, when deploying a model to a large number of users, we are usually more stringent regarding latency, as well as computational and storage capacity. Yet, simply using a smaller model does not take full advantage of the training capacity we usually possess.

A proposed solution to these seemingly opposing constrains is knowledge distillation. This approach involves training a large model, known as the teacher, and transferring its knowledge to a smaller model, called student, we want to deploy. We believe that the teacher is able to better extract the structure from the data. It learns to differentiate between large number of classes and then correctly predict the label when exposed to new data. Additionally, the trained model also assigns weights to all of the possible classes, which are then converted into probabilities using a softmax function. Even though these are often very small for the incorrect answers, they can still provide valuable information about how the larger model generalizes.

For example, an image of a horse will be correctly labeled by the teacher model with high probability close to 1. However, the model might also assign a small but nonzero probability that the image is a zebra. We argue that this probability will still be many times higher than the probability assigned to an unrelated class, such as a car.

Transferring this knowledge from the teacher to the student is done through distillation, where the student model is trained using the class probabilities produced by the teacher as soft targets. In the original paper, the distillation process is formulated as the minimization of the Kullback–Leibler (KL) divergence.

In this work, we propose enhancing the distillation process by replacing the KL divergence with Rényi divergence, which serves as its generalization, and introduces an additional hyperparameter α. We aim to formally define this new distillation framework, analyze the theoretical properties of Rényi-based distillation, and conduct experiments to evaluate the appropriateness of this approach.

1. Knowledge Distillation and Rényi Divergence

In this chapter, we formally define the notion of knowledge distillation as proposed in Hinton et al. [2015] and examine some of the theoretical results they presented. Next, we define Rényi divergence and examine some of its theoretical properties stated in van Erven and Harremoës [2012]. Some of these properties may justify our proposal to incorporate it into the distillation process via substitution for KL divergence.

1.1 Knowledge Distillation

Let us define model as a function that maps input data to output predictions on learned parameters

$$f_\theta : \mathcal{X} \to \mathcal{Y}, \tag{1.1}$$

where $\mathcal{X}$ is the input space, $\mathcal{Y}$ and θ represents the set of parameters of the model. In our case, where we assume a classification task with n classes, $\mathcal{Y} = \mathbb{R}^n$, so the model outputs a vector of n real-valued scores, referred to as logits. These logits are converted to probabilities using softmax function

$$p_c(x, \theta) = \frac{\exp z_c}{\sum_{k=1}^{n} \exp z_k}, \tag{1.2}$$

where $x \in \mathcal{X}$, $c \in (1, 2, \cdots, n)$, $f_\theta(x) = \mathbf{z}$ and $p(x)_c$ is the probability that x belongs to class c. We denote $P(x, \theta) = (p_1(x, \theta), p_2(x, \theta), \cdots, p_3(x, \theta))$.

We define a dataset $\mathcal{D}$ containing a total of N samples $\{(x_i, y_i)\}_{i=1}^{N}$, where for all i, $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$.

Additionally, since trained models produce peaky probability distributions, the approach is to soften these probabilities using temperature scaling. Thus we have

$$p_c^T(x, \theta) = \frac{\exp \frac{z_c}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}}, \tag{1.3}$$

where T is a hyperparameter. Popular choices for T according to Cho and Hariharan [2019] are 3, 4 and 5. During training, we apply this scaling to both the teacher and the student. By increasing T, we soften the probabilities, thus retaining inter-class similarities by driving the prediction away from 0 and 1.

Example. Following an example from the introduction, we can imagine a model f_θ that outputs logits for the classes horse, zebra, and car, equal to 5.4, 0.2, and -1.3, respectively. In Figure 1.1, we see the values in a bar chart, along with the computed probabilities using the softmax function, both without and with temperature scaling, the latter corresponding to $T = 4$.

Without temperature scaling, it is clear that the model is highly confident the input belongs to the class horse (> 0.99), while the probabilities for the remaining classes are essentially zero. However, as our goal is to improve the generalization of the model, we need to magnify the differences between classes with low probability. We can clearly see that by using temperature scaling, we achieve this effect.

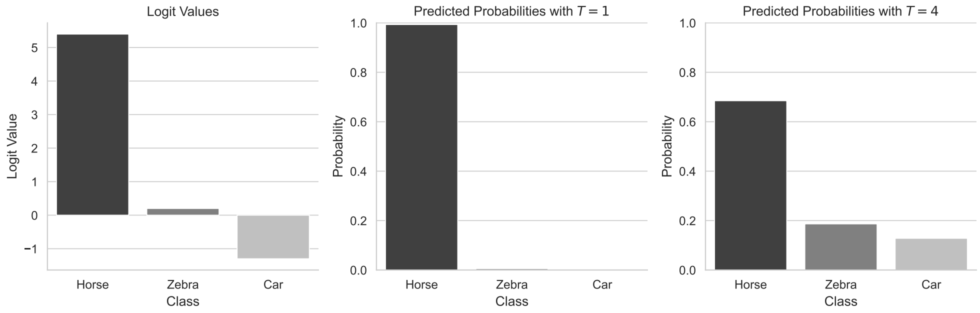

Figure 1.1: Example of temperature scaling.

Now that we know how to construct probabilities from the model output, we introduce a function to compare them.

Definition 1. *The Kullback–Leibler divergence of a probability distribution* $P = (p_1, \ldots, p_n)$ *from another distribution* $Q = (q_1, \ldots, q_n)$ *is*

$$D_{KL}(P\|Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}, \tag{1.4}$$

where we define $0 \log \frac{0}{0}$ *as* 0 *and* $x \log \frac{x}{0}$ *as* ∞ *for* $x > 0$.

We can decompose the KL divergence into two terms, as shown below

$$D_{\mathrm{KL}}(P\|Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{n} p_i \log p_i + \left(- \sum_{i=1}^{n} p_i \log q_i \right), \tag{1.5}$$

where the first term is the negative entropy of P, denoted as $-H(P)$, and the second term is the cross-entropy of P and Q, denoted as $H(P\|Q)$. Thus, we have

$$D_{\mathrm{KL}}(P\|Q) + H(P) = H(P\|Q).$$

And since $H(P)$ does not depend on Q, minimizing $D_{\mathrm{KL}}(P\|Q)$ with respect to Q is equivalent to minimizing $H(P\|Q)$. This, together with the fact that cross-entropy is easier to compute and interpret, especially when P represents one-hot labels, makes KL divergence often replaced by cross-entropy in machine learning.

Moreover, computing the derivative of KL divergence and cross-entropy with respect to Q yields the same result. But, in machine learning, we are interested in calculating the derivative with respect to the logits. Let's denote $q_i = \exp \frac{z_i}{T} / (\sum_{k=1}^{n} \exp \frac{z_k}{T})$, where z_i are the logits and T is the temperature. Now, we can compute the following

$$
\begin{aligned}
\frac{\partial D_{\mathrm{KL}}(P\|Q)}{\partial z_j} = \frac{\partial H(P\|Q)}{\partial z_j} &= -\frac{\partial}{\partial z_j}\sum_{i=1}^{n} p_i \log \frac{\exp \frac{z_i}{T}}{\sum_{k=1}^{n}\exp\frac{z_k}{T}}, && (1.6)\\
&= \left(\sum_{i=1}^{n} p_i\right)\frac{\partial}{\partial z_j}\log\sum_{k=1}^{n}\exp\frac{z_k}{T} - \frac{\partial}{\partial z_j}\sum_{i=1}^{n} p_i \frac{z_i}{T}, && (1.7)\\
&= \frac{1}{T}\frac{\exp z_j}{\sum_{k=1}^{n}\exp z_k} - \frac{p_j}{T} = \frac{1}{T}(q_j - p_j). && (1.8)
\end{aligned}
$$

Let f_t be a teacher model with parameters θ_t, and let f_s be a student model with parameters θ_s. We usually expect the teacher model to be structurally much larger than the student model, with significantly more parameters. During knowledge distillation, we aim to transfer knowledge from the teacher model, which is fully trained on $\mathcal{D}$, to the student model.

This transfer is achieved during the training of the student model, where the loss used is a linear combination of the standard cross-entropy loss with ground truth labels

$$
\begin{aligned}
\mathcal{L}_{\mathrm{CE}} &= \sum_{i=1}^{N} H(y_i\|P(x_i,\theta_s)), && (1.9)\\
&= -\sum_{i=1}^{N}\sum_{j=1}^{n} y_{i,j}\log p_j(x_i,\theta_s), && (1.10)
\end{aligned}
$$

and the KL divergence loss with the teacher's predictions

$$
\begin{aligned}
\mathcal{L}_{\mathrm{KL}} &= T^2\sum_{i=1}^{N} D_{\mathrm{KL}}(P^T(x_i,\theta_t)\|P^T(x_i,\theta_s)), && (1.11)\\
&= T^2\sum_{i=1}^{N}\sum_{j=1}^{n} p_j^T(x_i,\theta_t)\log\frac{p_j^T(x_i,\theta_t)}{p_j^T(x_i,\theta_s)}, && (1.12)
\end{aligned}
$$

where y_i is the one-hot ground-truth label, and T is a temperature. Thus, the total loss being optimized is

$$
\mathcal{L} = (1-\beta)\mathcal{L}_{\mathrm{CE}} + \beta\mathcal{L}_{\mathrm{KL}}, \qquad (1.13)
$$

where β is a hyperparameter that controls the balance between training on the true labels and training on the soft targets provided by the teacher model. A common choice for β is 0.9.

Lastly, as we look at the Equation 1.11 we can see the additional term T^2. This ensures that the relative contribution of hard and soft targets remains approximately the same. This result is dependent on an assumption given by Hinton et al. [2015], that both the logits of the student model z_j and the logits of the teacher model v_j are zero-meaned separately for each transfer case, such that $\sum_{k=1}^{n} z_k = \sum_{k=1}^{n} v_k = 0$. We have

$$
\frac{\partial H(P\|Q)}{\partial z_j} = \frac{1}{T}(q_j - p_j) = \frac{1}{T}\left(\frac{\exp\frac{z_j}{T}}{\sum_{k=1}^{n}\exp\frac{z_k}{T}} - \frac{\exp\frac{v_j}{T}}{\sum_{k=1}^{n}\exp\frac{v_k}{T}}\right). \qquad (1.14)
$$

Now, we approximate the exponential function using a Taylor polynomial for a temperature T that is high compared to the magnitude of the logits. We get

$$\frac{\partial H(P\|Q)}{\partial z_j} \approx \frac{1}{T}\left(\frac{1+\frac{z_j}{T}}{n+\sum_{k=1}^{n}\frac{z_k}{T}} - \frac{1+\frac{v_j}{T}}{n+\sum_{k=1}^{n}\frac{v_k}{T}}\right) = \frac{1}{nT^2}(z_i - v_i). \tag{1.15}$$

Thus, the gradient decreases quadratically as the temperature increases. To counter it we added T^2 in Equation 1.11. For lower temperature, Hinton et al. [2015] states that the distillation pays less attention to matching logits much more negative than average. This is advantageous, as they may be significantly noisier, given that the student model is not penalized for them during training. On the other hand, they might convey useful information about the knowledge acquired by the teacher. Based on empirical evidence, the authors claim that ignoring large negative logits has a positive effect, as intermediate temperatures yield the best results.

1.2 Rényi Divergence

Now that we have clearly defined knowledge distillation, we shift our focus to Rényi divergence, the alternative we propose to replace KL divergence. We start with the definition.

Definition 2. *The Rényi divergence of order α of a probability distribution $P = (p_1, \ldots, p_n)$ from another distribution $Q = (q_1, \ldots, q_n)$ is*

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha}, \tag{1.16}$$

where α is positive number distinct from 1, and we adopt the convention that $\frac{0}{0} = 0$ and $\frac{x}{0} = \infty$ for $x > 0$.

This definition of Rényi divergence assumes that probability distributions P and Q are discrete. For continuous spaces we can substitute the sum by Lebesgue integral.

Example. Let Q be a probability distribution, A be a set, such that $Q(A) > 0$. Define P as the conditional distribution of Q given A, i.e. $P(x) = Q(x|A) = \frac{Q(x)}{Q(A)}$, for $x \in A$. Now take the Rényi divergence of P from Q

$$\begin{aligned}
D_\alpha(P\|Q) &= \frac{1}{\alpha-1}\log\sum_{x\in A} P(x)^\alpha Q(x)^{1-\alpha}, &(1.17)\\
&= \frac{1}{\alpha-1}\log\sum_{x\in A}\left(\frac{Q(x)}{Q(A)}\right)^\alpha Q(x)^{1-\alpha}, &(1.18)\\
&= \frac{1}{\alpha-1}\log\sum_{x\in A}\frac{Q(x)}{Q(A)^\alpha}, &(1.19)\\
&= \frac{1}{\alpha-1}\log Q(A)^{1-\alpha} = -\log Q(A). &(1.20)
\end{aligned}$$

Here, we utilized the fact that $\sum_{x\in A} Q(x) = Q(A)$. In this particular example we observe that the factor $\frac{1}{\alpha-1}$ in the definition Rényi divergence has the effect that the derived value of $D_\alpha(P\|Q)$ does not depend on α.

In the introduction, we mentioned that Rényi divergence is a generalization of KL divergence. However, we cannot simply substitute a particular α to directly obtain the formula for KL divergence. To address this, we need to examine the limiting behavior.

There is a notion of extended orders, where we define Rényi divergence for values of α beyond the range specified in Definition 2, namely for $\alpha = 0$, $\alpha = 1$ and $\alpha = \infty$, as follows

$$\begin{aligned} D_0(P\|Q) &= \lim_{\alpha\to 0} D_\alpha(P\|Q), && (1.21)\\ D_1(P\|Q) &= \lim_{\alpha\to 1} D_\alpha(P\|Q), && (1.22)\\ D_\infty(P\|Q) &= \lim_{\alpha\to \infty} D_\alpha(P\|Q). && (1.23)\end{aligned}$$

Now, we need to calculate those limits. For $\alpha = 0$ we have

$$\begin{aligned} \lim_{\alpha\to 0} D_\alpha(P\|Q) &= \lim_{\alpha\to 0} \frac{1}{\alpha-1} \log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha}, && (1.24)\\ &= -\log \sum_{i=1}^{n} \lim_{\alpha\to 0} p_i^\alpha q_i^{1-\alpha}, && (1.25)\\ &= -\log \sum_{i=1}^{n} q_i \lim_{\alpha\to 0} p_i^\alpha. && (1.26)\end{aligned}$$

We can use the fact that $p_i^\alpha \to 1$ as $\alpha \to 0$ for $p_i > 0$. For $p_i = 0$, we get 0. Thus the Rényi divergence of order 0 is

$$D_0(P\|Q) = -\log \sum_{i=1}^{n} q_i \, \mathbb{1}\{p_i > 0\}, \tag{1.27}$$

where $\mathbb{1}$ is the indicator function. For $\alpha = 1$ the limit

$$\lim_{\alpha\to 0} \frac{1}{\alpha-1} \log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha} \tag{1.28}$$

is of an indeterminate form $\dfrac{0}{0}$, allowing us to apply L'Hopital's Rule.

$$\begin{aligned} \lim_{\alpha\to 0} \frac{1}{\alpha-1} \log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha} &= \lim_{\alpha\to 0} \frac{\sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha} \log p_i - p_i^\alpha q_i^{1-\alpha} \log q_i}{\sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha}}, && (1.29)\\ &= \frac{\sum_{i=1}^{n} p_i \log p_i - p_i \log q_i}{\sum_{i=1}^{n} p_i}, && (1.30)\\ &= \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}. && (1.31)\end{aligned}$$

This yields that $D_1(P\|Q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$, which is also equal to $D_{\mathrm{KL}}(P\|Q)$. Thus, we have proven that Rényi divergence generalizes KL divergence and is equal to KL divergence when $\alpha = 1$. Lastly, for $\alpha = \infty$, we compute the limit

$$\lim_{\alpha \to \infty} \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} p_i^{\alpha} q_i^{1-\alpha}. \tag{1.32}$$

We know that the term $\max_i q_i \left(\frac{p_i}{q_i}\right)^{\alpha}$ dominates $\sum_{i=1}^n p_i^{\alpha} q_i^{1-\alpha}$ as $\alpha \to \infty$. So now we have

$$\frac{1}{\alpha - 1} \log \max_i q_i \left(\frac{p_i}{q_i}\right)^{\alpha} \to \max_i \log \frac{p_i}{q_i}, \tag{1.33}$$

giving us the result $D_\infty(P\|Q) = \max_i \log \frac{p_i}{q_i}$.

In van Erven and Harremoës [2012] we can find the corresponding theorems and proves for continuous probability distributions.

2. Title of the second chapter

2.1 Title of the first subchapter of the second chapter

2.2 Title of the second subchapter of the second chapter

Conclusion

Bibliography

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. *CoRR*, abs/1910.01348, 2019. URL `http://arxiv.org/abs/1910.01348`.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *CoRR*, abs/1206.2459, 2012. URL `http://arxiv.org/abs/1206.2459`.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment