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 Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague

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 kybernetiky, Fakulta elektrotechnická, České vysoké učení technické v Praze

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. Rényi Divergence and Knowledge Distillation

In this chapter, we begin by examining the concepts of entropy, cross-entropy, and divergence. In particular, we define Rényi divergence, establish its connection to KL divergence, and inspect some of its theoretical properties stated in van Erven and Harremoës [2012]. In the second part, we formally define the notion of knowledge distillation, as proposed in Hinton et al. [2015], and examine some of the theoretical results they presented. Furthermore, we analyze how these results change when incorporating Rényi divergence into the distillation process.

1.1 KL Divergence and Rényi Divergence

The concept of entropy, as the amount of uncertainty regarding the outcome of an experiment, was introduced by Shannon [1948].

Definition 1. *The entropy of a probability distribution $P = (p_1, \ldots, p_n)$ is given by*

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i.$$

Example. Let P be the probability distribution of a fair coin toss, i.e., $P = (\frac{1}{2}, \frac{1}{2})$. The entropy $H(P)$ is approximately 0.693. Next, let Q represent the probability distribution of a slightly unfair coin toss, i.e., $Q = (\frac{4}{10}, \frac{6}{10})$. Here, the entropy $H(Q)$ is smaller than $H(P)$, approximately 0.673. In other words, we are less uncertain about the outcome of the unfair coin toss than about the fair coin toss.

To determine the similarity between two probability distributions, we cannot simply subtract their entropies. For example, the entropy of $P_1 = (\frac{4}{10}, \frac{6}{10})$ is the same as the entropy of $P_2 = (\frac{6}{10}, \frac{4}{10})$, yet they represent different distributions. Therefore, we use the concept of divergence, as proposed by Kullback and Leibler [1951], which is also closely related to cross-entropy, a measure that is more prominently used in machine learning applications.

Definition 2. *The cross-entropy of a probability distribution $P = (p_1, \ldots, p_n)$ from another distribution $Q = (q_1, \ldots, q_n)$ is given by*

$$H(P, Q) = -\sum_{i=1}^{n} p_i \log q_i.$$

Example. Let P and Q be the probability distributions as described in the example above, i.e., $P = (\frac{1}{2}, \frac{1}{2})$ and $Q = (\frac{4}{10}, \frac{6}{10})$. The cross-entropy of P from Q is $H(P, Q) \approx 0.714$. Moreover, cross-entropy is not symmetric, thus $H(P, Q) \neq H(Q, P) \approx 0.693$.

Definition 3. *The Kullback–Leibler divergence of a probability distribution* $P = (p_1, \ldots, p_n)$ *from another distribution* $Q = (q_1, \ldots, q_n)$ *is given by*

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

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

$$\begin{aligned} D_{\mathrm{KL}}(P \| Q) &= \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \\ &= \sum_{i=1}^{n} p_i \log p_i + \left(- \sum_{i=1}^{n} p_i \log q_i \right), \\ &= -H(P) + H(P, Q) \end{aligned} \tag{1.1}$$

and we observe that cross-entropy can be decomposed into entropy and KL divergence.

Example. Let P and Q be probability distributions, such as $P = (\frac{1}{2}, \frac{1}{2})$ and $Q = (\frac{4}{10}, \frac{6}{10})$. From the previous examples, we know that $H(P) \approx 0.693$ and $H(P, Q) \approx 0.714$. Using Equation 1.1, we can calculate the KL divergence of P from Q as $D_{\mathrm{KL}}(P \| Q) = -H(P) + H(P, Q) \approx 0.021$.

Kullback–Leibler divergence was later generalized by Rényi [1961]. We begin with its definition.

Definition 4. *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 given by*

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

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 (see van Erven and Harremoës [2012]). Now, we present an example that motivated introduction of the normalization term $\frac{1}{\alpha-1}$ in the definition.

Example. Let Q be a probability distribution and 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

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

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.

Definition 4 was formulated for orders $\alpha \in (0,1)\cup(1,\infty)$. We now show that the limits on the borders of the domain for α exist and therefore Rényi divergence can be naturally extended to the cases $\alpha = 0, 1, \infty$. That is, we inspect the limits

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

where P and Q are discrete distributions on $(1,\cdots,n)$. For $\alpha = 0$, we have

$$
\begin{aligned}
\lim_{\alpha\to 0+} D_\alpha(P\|Q) &= \lim_{\alpha\to 0+}\frac{1}{\alpha-1}\log\sum_{i=1}^n p_i^\alpha q_i^{1-\alpha},\\
&= -\log\sum_{i=1}^n \lim_{\alpha\to 0+} p_i^\alpha q_i^{1-\alpha},\\
&= -\log\sum_{i=1}^n q_i \lim_{\alpha\to 0+} p_i^\alpha\\
&= -\log\sum_{i=1}^n q_i\,\mathbb{1}\{p_i > 0\},
\end{aligned}
\tag{1.2}
$$

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

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

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

$$
\begin{aligned}
\lim_{\alpha\to 1}\frac{1}{\alpha-1}\log\sum_{i=1}^{n}p_i^{\alpha}q_i^{1-\alpha} &= \lim_{\alpha\to,}\frac{\sum_{i=1}^{n}p_i^{\alpha}q_i^{1-\alpha}\log p_i - p_i^{\alpha}q_i^{1-\alpha}\log q_i}{\sum_{i=1}^{n}p_i^{\alpha}q_i^{1-\alpha}}, \\
&= \frac{\sum_{i=1}^{n}p_i\log p_i - p_i\log q_i}{\sum_{i=1}^{n}p_i}, \\
&= \sum_{i=1}^{n}p_i\log\frac{p_i}{q_i}.
\end{aligned} \tag{1.3}
$$

Lastly, for $\alpha=\infty$, we define the summation term of the limit as $Z(\alpha)=\sum_{i=1}^{n}p_i^{\alpha}q_i^{1-\alpha}$. Additionally, define $M=\max_i\frac{p_i}{q_i}$, and let I be the set of indices that achieve this maximum. For large α, the terms where $\frac{p_i}{q_i}$ is the largest will dominate, leading to the asymptotic approximation

$$
Z(\alpha)\approx M^{\alpha}\sum_{i\in I}q_i.
$$

Taking the logarithm and dividing by $\alpha-1$

$$
\frac{1}{\alpha-1}\log Z(\alpha)\approx\frac{\alpha logM+\log\sum_{i\in I}q_i}{\alpha-1}.
$$

Taking the limit as $\alpha\to\infty$

$$
\begin{aligned}
\lim_{\alpha\to\infty}\frac{1}{\alpha-1}\log\sum_{i=1}^{n}p_i^{\alpha}q_i^{1-\alpha} &= \log M, \\
&= \max_i\log\frac{p_i}{q_i}.
\end{aligned} \tag{1.4}
$$

The limits 1.2, 1.3 , 1.4 allow us to define the Rényi divergences

$$
\begin{aligned}
D_0(P\|Q) &= -\log\sum_{i=1}^{n}q_i\,\mathbb{1}\{p_i>0\}, \\
D_1(P\|Q) &= \sum_{i=1}^{n}p_i\log\frac{p_i}{q_i}, \\
D_\infty(P\|Q) &= \max_i\log\frac{p_i}{q_i}.
\end{aligned}
$$

Comparing to Definition 3, we see that

$$
D_1(P\|Q)=D_{\mathrm{KL}}(P\|Q),
$$

and Rényi divergence indeed generalizes KL divergence.

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

Another important case of Rényi divergence is for $\alpha=\frac{1}{2}$. For only this value the Rényi divergence is symmetric, i.e., $D_{1/2}(P\|Q)=D_{1/2}(Q\|P)$. Even with this additional property, it still does not satisfy the definition of a metric, as a property of triangle inequality does not hold. However, Rényi divergence of order $\frac{1}{2}$ can be rewritten as a function of the squared Hellinger distance, which is defined for discrete probability distributions P and Q as follows

$$H^2(P\|Q) = \frac{1}{2}\sum_{i=1}^{n}(p_i^{\frac{1}{2}} - q_i^{\frac{1}{2}})^2.$$

Rewriting it, we get

$$\frac{1}{2}\sum_{i=1}^{n}(p_i^{\frac{1}{2}} - q_i^{\frac{1}{2}})^2 = \frac{1}{2}\left(\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} q_i - 2\sum_{i=1}^{n} p_i^{\frac{1}{2}} q_i^{\frac{1}{2}}\right) = 1 - \sum_{i=1}^{n} p_i^{\frac{1}{2}} q_i^{\frac{1}{2}},$$

which we can use in the definition of Rényi divergence of order $\frac{1}{2}$ to express it in terms of the Hellinger distance

$$D_{1/2}(P\|Q) = \frac{1}{\frac{1}{2}-1}\log\sum_{i=1}^{n} p_i^{\frac{1}{2}} q_i^{1-\frac{1}{2}} = -2\log(1 - H^2(P\|Q)).$$

We can also establish a connection between Rényi divergence of order α and $1-\alpha$ for $0 < \alpha < 1$.

$$\begin{aligned} D_{1-\alpha}(P\|Q) &= \frac{1}{-\alpha}\log\sum_{i=1}^{n} p_i^{1-\alpha} q_i^{\alpha}, \\ &= \frac{1-\alpha}{\alpha}\left(\frac{\alpha}{1-\alpha}\frac{1}{-\alpha}\log\sum_{i=1}^{n} q_i^{\alpha} p_i^{1-\alpha}\right), \\ &= \frac{1-\alpha}{\alpha} D_{\alpha}(Q\|P). \end{aligned}$$

This relationship can be rewritten as in van Erven and Harremoës [2012] to obtain

$$D_{\alpha}(P\|Q) = \frac{\alpha}{1-\alpha} D_{1-\alpha}(Q\|P).$$

1.2 Knowledge Distillation

Let us define a machine learning model as a function that maps input data to output predictions

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

where $\mathcal{X}$ is the input space, $\mathcal{Y}$ is the output space and θ represents the set of parameters of the model. In our case, as input, the model receives images from $\mathcal{X}$, where each image is represented as a tensor in $\mathbb{R}^{h\times w\times c}$. We assume a classification task with n classes, i.e., $\mathcal{Y} = \mathbb{R}^n$, so the model outputs a vector of n real-valued scores, referred to as logits. These logits are converted to probabilities using softmax function

$$\tilde{p}_c(x,\theta) = \frac{\exp z_c}{\sum_{k=1}^{n}\exp z_k}, \qquad c = 1, 2, \cdots, n,$$

for $x \in \mathcal{X}$, $z_k = f_{\theta}(x)_k$ for all $k \in \{1, 2, \cdots, n\}$. Then $\tilde{p}_c(x,\theta)$ represents the probability that x belongs to class c. Denote $\tilde{P}(x,\theta) = (\tilde{p}_1(x,\theta), \tilde{p}_2(x,\theta), \cdots, \tilde{p}_n(x,\theta))$.

Additionally, a hyperparameter T, called temperature, is introduced to control the entropy of the output distribution. That is we set for $T > 0$

$$p_c(x, \theta) = \frac{\exp \frac{z_c}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}}, \qquad c = 1, 2, \cdots, n.$$

Denote $P(x, \theta) = (p_1(x, \theta), p_2(x, \theta), \cdots, p_n(x, \theta))$. The process in called temperature scaling and popular choices for T according to Cho and Hariharan [2019] are 3, 4 and 5.

Example. Following the example from the introduction, suppose a model f_θ that for given input yields 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 logit 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, the model is highly confident that the input belongs to the class horse (> 0.99), while the probabilities for the remaining classes are essentially zero. We observe that the effect of the temperature scaling is that the model is less confident about the true label while the order of the class probabilities in maintained.

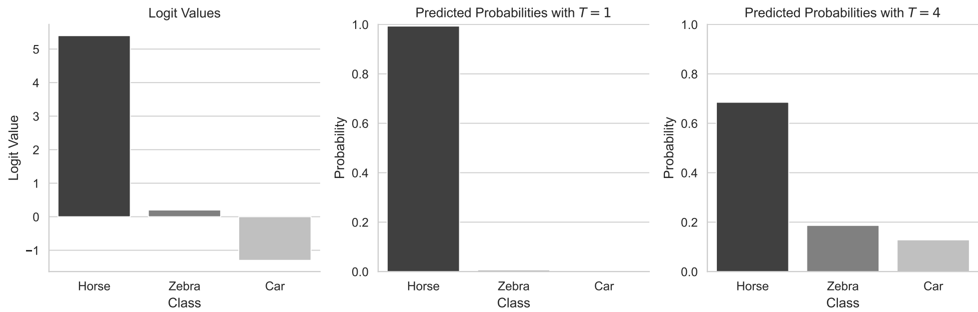

Figure 1.1: Example of temperature scaling.

Let $\hat{\mathcal{D}} = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{\hat{N}}$ be a training dataset, containing a total of $\hat{N}$ samples, where $(\hat{x}_i, \hat{y}_i) \in \mathcal{X} \times \mathcal{Y}$ for all $i \in \{1, 2, \cdots, \hat{N}\}$. Let Similarly, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a transfer dataset, containing a total of N samples, where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for all $i \in \{1, 2, \cdots, N\}$. In many cases, we have $\hat{\mathcal{D}} = \mathcal{D}$.

Definition 5. *Knowledge distillation is a model compression technique where a smaller student model f_s with parameters θ_s is trained to mimic a larger teacher model f_t with parameters θ_t, which has been pre-trained on a dataset $\hat{\mathcal{D}}$. The training process utilizes a transfer dataset $\mathcal{D}$ and optimizes the loss function of the form*

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

where $\mathcal{L}_{CE}$ is the standard cross-entropy loss with ground truth labels

$$\begin{aligned} \mathcal{L}_{CE} &= \sum_{i=1}^{N} H(y_i \| \tilde{P}(x_i, \theta_s)), \\ &= -\sum_{i=1}^{N} \sum_{j=1}^{n} y_{i,j} \log \tilde{p}_j(x_i, \theta_s), \end{aligned} \tag{1.5}$$

and $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence loss with teacher's predictions

$$\begin{aligned}\mathcal{L}_{KL} &= T^2 \sum_{i=1}^{N} D_{KL}(P(x_i,\theta_t) \| P(x_i,\theta_s)),\\ &= T^2 \sum_{i=1}^{N}\sum_{j=1}^{n} p_j(x_i,\theta_t)\log\frac{p_j(x_i,\theta_t)}{p_j(x_i,\theta_s)},\end{aligned} \tag{1.6}$$

y_i is the one-hot ground-truth label, and T and β are hyperparameters.

The hyperparameter T in Definition 5 denotes temperature. During training, we apply temperature scaling to both the teacher and the student in Equation 1.6. By increasing T, we soften the probabilities, thus retaining inter-class similarities by driving the predictions away from 0 and 1. The second hyperparameter, β, controls the balance between training on the truth labels and training on the soft targets provided by the teacher model. A common choice for β is 0.9 (see Cho and Hariharan [2019]).

We observe that, unlike in Equation 1.5, which is simply the sum of the cross-entropies, in Equation 1.6, the total loss also includes the term T^2. To understand the origin of this term we firstly need to calculate the derivative of KL divergence $D_{\mathrm{KL}}(P\|Q)$ with respect to the logits of $Q = (q_1, q_2, \cdots, q_n)$, i.e., $q_i = \exp\frac{z_i}{T}/(\sum_{k=1}^{n}\exp\frac{z_k}{T})$ for all $i \in \{1, 2, \cdots, n\}$, where z_i are the logits and T is a temperature.

Using Equation 1.1, we observe that since $H(P)$ does not depend on Q, the derivative of $D_{\mathrm{KL}}(P\|Q)$ with respect to Q is equivalent to the derivative of $H(P\|Q)$. This, together with the fact that cross-entropy is easier to compute, especially when P represents one-hot labels, makes KL divergence often replaced by cross-entropy in machine learning application. Now, we compute the following

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

From this result and 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 $\sum_{k=1}^{n} z_k = \sum_{k=1}^{n} v_k = 0$. For $q_i = \exp\frac{z_i}{T}/(\sum_{k=1}^{n}\exp\frac{z_k}{T})$ and $p_i = \exp\frac{v_i}{T}/(\sum_{k=1}^{n}\exp\frac{v_k}{T})$ for all $i \in \{1, 2, \cdots, n\}$ we have

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

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

$$\frac{\partial H(P\|Q)}{\partial z_j} \approx \frac{1}{T}\left(\frac{1+\frac{z_j}{T}}{n+\sum_{k=1}^{n}\frac{z_k}{T}} - \frac{1+\frac{v_j}{T}}{n+\sum_{k=1}^{n}\frac{v_k}{T}}\right) = \frac{1}{nT^2}(z_j - v_j). \tag{1.7}$$

Thus, the gradient decreases proportionally to $\frac{1}{T^2}$ as the temperature T increases. By incorporating the term T^2 into Equation 1.6, we ensure that the relative contribution of hard and soft targets remains approximately the same.

For lower temperature, where the approximation by the Taylor polynomial is very inaccurate,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.

Incorporating Rényi divergence into the knowledge distillation is done by replacing the KL divergence loss $\mathcal{L}_{\mathrm{KL}}$ by the Rényi divergence loss $\mathcal{L}_\alpha$, where $\alpha \geq 0$ is the order of the Rényi divergence. Similarly to the $\mathcal{L}_{\mathrm{KL}}$, the Rényi divergence loss must also be the product of two terms, since $\mathcal{L}_{\mathrm{KL}} = \mathcal{L}_1$. In the summation term, we only replace $D_{\mathrm{KL}}(P(x_i, \theta_t) \| P(x_i, \theta_s)))$ by $D_\alpha(P(x_i, \theta_t) \| P(x_i, \theta_s))$.

What remains less evident is how the T^2 term is affected. Firstly, we compute the derivative of $D_\alpha(P\|Q)$ with respect to the logits of Q, given that $q_i = \exp \frac{z_i}{T} / (\sum_{k=1}^n \exp \frac{z_k}{T})$, where T is a temperature. We have

$$\frac{\partial D_\alpha(P\|Q)}{\partial z_j} = \frac{\partial}{\partial z_j} \frac{1}{\alpha - 1} \log \sum_{i=1}^n p_i^\alpha \left(\frac{\exp \frac{z_i}{T}}{\sum_{k=1}^n \exp \frac{z_k}{T}} \right)^{1-\alpha},$$

and denote $Z = \sum_{i=1}^n p_i^\alpha \left(\frac{\exp \frac{z_i}{T}}{\sum_{k=1}^n \exp \frac{z_k}{T}} \right)^{1-\alpha}$. From that we use the chain rule

$$\frac{\partial D_\alpha(P\|Q)}{\partial z_j} = \frac{\partial}{\partial z_j} \frac{1}{\alpha - 1} \log Z = \frac{1}{\alpha - 1} \frac{1}{Z} \frac{\partial Z}{\partial z_j}.$$

Now we need to calculate $\frac{\partial Z}{\partial z_j}$.

$$\begin{aligned}
\frac{\partial Z}{\partial z_j} =& \frac{\partial}{\partial z_j} \sum_{i=1}^{n} p_i^{\alpha} \left(\frac{\exp \frac{z_i}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right)^{1-\alpha}, \\
=& \frac{\partial}{\partial z_j} p_j^{\alpha} \left(\frac{\exp \frac{z_j}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right)^{1-\alpha} + \frac{\partial}{\partial z_j} \sum_{i \neq j}^{n} p_i^{\alpha} \left(\frac{\exp \frac{z_i}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right)^{1-\alpha}, \\
=& p_j^{\alpha} (1-\alpha) \left(\frac{\exp \frac{z_j}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right)^{-\alpha} \frac{\frac{1}{T} \sum_{k=1}^{n} \exp \frac{z_k}{T} \exp \frac{z_j}{T} - \frac{1}{T} \exp \frac{z_j}{T} \exp \frac{z_j}{T}}{(\sum_{k=1}^{n} \exp \frac{z_k}{T})^2} \\
& - \sum_{i \neq j}^{n} p_i^{\alpha} \left(\frac{\exp \frac{z_i}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right)^{-\alpha} (1-\alpha) \frac{\exp \frac{z_i}{T}}{(\sum_{k=1}^{n} \exp \frac{z_k}{T})^2} \exp \frac{z_j}{T} \frac{1}{T}, \\
=& \frac{1-\alpha}{T} \left[p_j^{\alpha} \left(\frac{\exp \frac{z_j}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right)^{1-\alpha} \frac{\sum_{k=1}^{n} \exp \frac{z_k}{T} - \exp \frac{z_j}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right. \\
& \left. - \sum_{i \neq j}^{n} p_i^{\alpha} \left(\frac{\exp \frac{z_i}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right)^{1-\alpha} \frac{\exp \frac{z_j}{T}}{\sum_{k=1}^{n} \exp \frac{z_k}{T}} \right]. \\
=& \frac{1-\alpha}{T} \left[p_j^{\alpha} q_j^{1-\alpha} (1-q_j) - \sum_{i \neq j}^{n} p_i^{\alpha} q_i^{1-\alpha} q_j \right], \\
=& \frac{1-\alpha}{T} \left[p_j^{\alpha} q_j^{1-\alpha} - q_j \sum_{i=1}^{n} p_i^{\alpha} q_i^{1-\alpha} \right], \\
=& \frac{1-\alpha}{T} (p_j^{\alpha} q_j^{1-\alpha} - q_j Z).
\end{aligned}$$

Now, inserting $\frac{\partial Z}{\partial z_j}$ into the original equation, we obtain

$$\frac{\partial D_{\alpha}(P \| Q)}{\partial z_j} = \frac{1}{\alpha - 1} \frac{\frac{1-\alpha}{T}(p_j^{\alpha} q_j^{1-\alpha} - q_j Z)}{Z} = \frac{q_j Z - p_j^{\alpha} q_j^{1-\alpha}}{TZ}.$$

We can also substitute the original expression for Z and simplify the result to derive

$$\frac{\partial D_{\alpha}(P \| Q)}{\partial z_j} = \frac{1}{T} \left(q_j - \frac{p_j^{\alpha} q_j^{1-\alpha}}{\sum_{i=1}^{n} p_i^{\alpha} q_i^{1-\alpha}} \right), \tag{1.8}$$

where for $\alpha = 1$, we arrive at the same result as for KL divergence.

If the distribution P represents one-hot labels, the derivative simplifies to $\frac{q_j}{T}$ if $p_j = 0$, or to $\frac{q_j - 1}{T}$ if $p_j = 1$. This holds for any choice of α, which is why we do not modify the $\mathcal{L}_{\mathrm{CE}}$ term in knowledge distillation.

From the result in Equation 1.8, we must derive an analogous expression to Equation 1.7. Let z_j be the logits of the student model and let v_j be the logits of the teacher model. Here, we also use the assumption from Hinton et al. [2015], that the logits are zero-meaned separately for each transfer case, meaning that $\sum_{k=1}^{n} z_k = \sum_{k=1}^{n} v_k = 0$.

Moreover, we approximate the exponential function using a Taylor polynomial, but here, unlike before, the temperature must be high compared to both the magnitude of the logits and the value of α. We derive

$$
\begin{aligned}
\frac{\partial D_\alpha(P\|Q)}{\partial z_j} =& \frac{1}{T}\left[\frac{\exp\frac{z_j}{T}}{\sum_{k=1}^n \exp\frac{z_k}{T}} - \left(\frac{\exp\frac{v_j}{T}}{\sum_{k=1}^n \exp\frac{v_k}{T}}\right)^\alpha \left(\frac{\exp\frac{z_j}{T}}{\sum_{k=1}^n \exp\frac{z_k}{T}}\right)^{1-\alpha}\right.\\
&\left.\left(\sum_{i=1}^n \left(\frac{\exp\frac{v_i}{T}}{\sum_{k=1}^n \exp\frac{v_k}{T}}\right)^\alpha \left(\frac{\exp\frac{z_i}{T}}{\sum_{k=1}^n \exp\frac{z_k}{T}}\right)^{1-\alpha}\right)^{-1}\right],\\
\approx& \frac{1}{T}\left[\frac{1+\frac{z_j}{T}}{n+\sum_{k=1}^n \frac{z_k}{T}} - \frac{1+\frac{\alpha v_j}{T}}{\left(n+\sum_{k=1}^n \frac{v_k}{T}\right)^\alpha}\frac{1+\frac{(1-\alpha)z_j}{T}}{\left(n+\sum_{k=1}^n \frac{z_k}{T}\right)^{1-\alpha}}\right.\\
&\left.\left(\sum_{i=1}^n \frac{1+\frac{\alpha v_i}{T}}{\left(n+\sum_{k=1}^n \frac{v_k}{T}\right)^\alpha}\frac{1+\frac{(1-\alpha)z_i}{T}}{\left(n+\sum_{k=1}^n \frac{z_k}{T}\right)^{1-\alpha}}\right)^{-1}\right],\\
=& \frac{1}{T}\left[\frac{1+\frac{z_j}{T}}{n} - \frac{1+\frac{\alpha v_j}{T}+\frac{(1-\alpha)z_j}{T}+\frac{\alpha(1-\alpha)v_j z_j}{T^2}}{n}\right.\\
&\left.\left(\frac{1}{n}\sum_{i=1}^n 1+\frac{\alpha v_i}{T}+\frac{(1-\alpha)z_i}{T}+\frac{\alpha(1-\alpha)v_i z_i}{T^2}\right)^{-1}\right].
\end{aligned}
$$

When using the approximation of the exponential function by the first Taylor polynomial, we assume that the higher-order terms are negligible. Particularly, this means that $\alpha^2 v_j^2 = o(T^2)$ and $(1-\alpha)^2 z_j^2 = o(T^2)$. From this we get αv_j and $(1-\alpha)z_j$ are $o(T)$. Thus, their product is $o(T^2)$, which means that $\frac{\alpha(1-\alpha)v_j z_j}{T^2} \approx 0$.

This, together with the previously mentioned assumption of zero-meaned logits, allows us to further simplify the formula.

$$
\frac{\partial D_\alpha(P\|Q)}{\partial z_j} \approx \frac{1}{T}\left[\frac{1+\frac{z_j}{T}}{n} - \frac{1+\frac{\alpha v_j}{T}+\frac{(1-\alpha)z_j}{T}}{n}\left(\frac{n}{n}\right)^{-1}\right] = \frac{\alpha}{nT^2}(z_j - v_j).
$$

This formula is similar to that of KL divergence (1.7), except that it is multiplied by α. Thus the Rényi divergence loss is of this form

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

(Lower semi-continuous property here later)

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.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

Alfréd Rényi. On measures of entropy and information. 1961. URL `https://api.semanticscholar.org/CorpusID:123056571`.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948. URL `http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`.

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