Adaptive Multi-head Contrastive Learning –Appendix–

Lei Wang[⋆,1,2], Piotr Koniusz[2,1], Tom Gedeon[3], and Liang Zheng[1]

[1]Australian National University [2]Data61/CSIRO [3]Curtin University
{lei.w, liang.zheng}@anu.edu.au, piotr.koniusz@data61.csiro.au, tom.gedeon@curtin.edu.au

A Deriving the AMCL loss from MLE

This section details the derivation of our loss function based on the maximum likelihood estimation (MLE) over head-wise posterior distributions of positive samples given observations. We show that our derivation is connected to an m-estimator [27] whose log-likelihood employs Normal distributions a.k.a. Welsch functions that are known to model the observation noise via the heteroscedastic aleatoric uncertainty [28,30,38]. Our adaptive temperature captures such an uncertainty. Tuning constant τ was shown before to help learn good contrastive representations [9,21]. [44] also demonstrated that temperature τ controls the strength of penalties on the hard negative samples and established its relationship with uniformity, illustrating that a well-chosen τ can strike a balance between the alignment and uniformity properties of contrastive loss. [34] has shown that in place of constant temperature, a cosine schedule can improve learning–a seemingly minor modification with large impact on the learned embedding space.

For ℓ_2 normalized vectors, the relationship between squared Euclidean distance $\|\cdot\|_2^2$ and cosine similarity measure is: $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 = 2 - \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)$. The Normal distribution $\mathcal{N}$ relies on the squared Euclidean distance. To derive our multi-head NT-Xent loss, consider the following maximum likelihood estimation *w.r.t.* parameters given as $\mathcal{P} = \big\{\boldsymbol{\theta}, \{\tau_i^{c+}\}_{c=1}^C, \{\{\tau_{in}^{c-}\}_{n=1}^N\}_{c=1}^C\big\}$ and $\beta = 1$:

$$\mathcal{P}^* = \arg\max_{\mathcal{P}} \prod_{c=1}^{C} \frac{\mathcal{N}\big(2 - 2\text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_i^{c+}); \tau_i^{c+}\big)}{\sum_{n=1}^{N} \mathcal{N}\big(2 - 2\text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_{in}^{c-}); \tau_{in}^{c-}\big)} \tag{6}$$

$$\begin{aligned}
&= \arg\min_{\mathcal{P}} \sum_{c=1}^{C} \Big(-\log \mathcal{N}\big(2 - 2\text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_i^{c+}); \tau_i^{c+}\big) + \log \sum_{n=1}^{N} \mathcal{N}\big(2 - 2\text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_{in}^{c-}); \tau_{in}^{c-}\big)\Big)\\
&= \arg\min_{\mathcal{P}} \sum_{c=1}^{C} \Big(-\frac{1}{\tau_i^{c+}} \text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_i^{c+}) + \beta \Omega(\tau_i^{c+})\\
&+ \log \sum_{n=1}^{N} \frac{1}{(2\pi)^{d'/2} (\tau_{in}^{c-})^{d'/2}} \exp\Big(\frac{1}{\tau_{in}^{c-}}\big(\text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_{in}^{c-}) - 1\big)\Big).
\end{aligned} \tag{7}$$

[⋆] Corresponding author.

In Eq. (7), we simply use expansion:

$$-\log\Big(\frac{1}{(2\pi)^{d'/2}(\sigma^2)^{d'/2}}\exp\Big(-\frac{2-2\mathbf{s}}{2\sigma^2}\Big)\Big)=d'/2\log(2\pi)+(d'/2)\log(\sigma^2)+1/\sigma^2-\mathbf{s}/\sigma^2, \tag{8}$$

where variance $\sigma^2=\tau$. We drop the constant (no impact on optimization) and are left with $-\mathbf{s}/\tau$ and $\Omega(\tau)=(d'/2)\log(\tau)+1/\tau$. We apply approximation in Eq. (5) to Eq. (7) (rightmost part) and readily obtain Eq. (2.1). To derive multi-head InfoNCE loss, we solve a slightly modified problem:

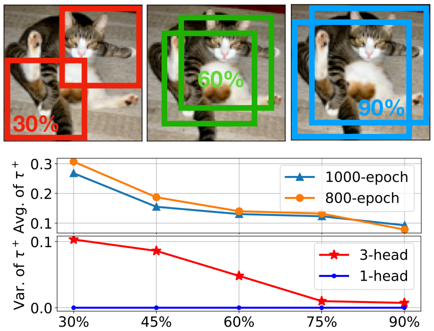
$$\begin{aligned}\mathcal{P}^*&=\arg\max_{\mathcal{P}}\prod_{c=1}^{C}\frac{\mathcal{N}\big(2-2\text{sim}(\boldsymbol{z}_i^c,\boldsymbol{z}_i^{c+});\tau_i^{c+}\big)}{\mathcal{N}\big(2-2\text{sim}(\boldsymbol{z}_i^c,\boldsymbol{z}_i^{c+});\tau_i^{c+}\big)+\sum_{n=1}^{N}\mathcal{N}\big(2-2\text{sim}(\boldsymbol{z}_i^c,\boldsymbol{z}_{in}^{c-});\tau_{in}^{c-}\big)} \qquad (9)\\ &=\arg\max_{\mathcal{P}}\prod_{c=1}^{C}\frac{\mathcal{N}\big(2-2\text{sim}(\boldsymbol{z}_i^c,\boldsymbol{z}_i^{c+});\tau_i^{c+}\big)}{\sum_{n=1}^{N+1}\mathcal{N}\big(2-2\text{sim}(\boldsymbol{z}_i^c,\boldsymbol{z}_{in}^{c\pm});\tau_{in}^{c\pm}\big)}=\arg\max_{\mathcal{P}}\prod_{c=1}^{C}p(\boldsymbol{z}_{ic}^{c+}|\boldsymbol{z}_i^c)\\ &=\arg\max_{\mathcal{P}}\prod_{c=1}^{C}\frac{p(\boldsymbol{z}_i^c|\boldsymbol{z}_{ic}^{c+})p(\boldsymbol{z}_{ic}^{c+})}{p(\boldsymbol{z}_i^c)},\end{aligned}$$

where $p(\boldsymbol{z}_i^c)=\sum_{n=1}^{N+1}\mathcal{N}\big(2-2\text{sim}(\boldsymbol{z}_i^c,\boldsymbol{z}_{in}^{c\pm});\tau_{in}^{c\pm}\big)$, $p(\boldsymbol{z}_{ic}^{c+})$ is a constant, *e.g.*, 1, and $p(\boldsymbol{z}_i^c|\boldsymbol{z}_{ic}^{c+})=\mathcal{N}\big(2-2\text{sim}(\boldsymbol{z}_i^c,\boldsymbol{z}_i^{c+});\tau_i^{c+}\big)$. Thus, the ratio of Gaussians in Eq. (9) can be interpreted as maximizing head-wise posterior distributions of positive samples given observations.

Connecting temperature to uncertainty. Eq. (6) uses the variance τ of the distribution of pair-wise distances. Eq. (6) derives Eq. (7), where τ weighs the similarity, making it effectively the temperature. Because variance is usually treated as uncertainty [46, 52], we build natural correspondence between uncertainty and temperature. As we derive our multi-head losses (*e.g.*, InfoNCE) from the MLE, we optimize this problem over network parameters and temperature (parametrized by an MLP). The temperature is tied with Welsch functions (Gaussians) in Eq. (6) and (9) whose radii are known to determine their influence range (tolerance to outliers).

B More Discussions

Criterion to define positive pair. Positive pairs that form two views are generated by several augmentations of an image. Fig. 1 (pair indicated by green dot) in the main paper shows different crops of a sheep (1st pair) and car colors/shapes (4th pair). For stronger augmentations (*e.g.*, low overlap of two crops) the noise effect on the contrastive loss becomes stronger (*e.g.*, disjoint positive box of cat's leg may be shared between different heads). Thus, our positive-pair temperature obtained from the aleatoric uncertainty learner downweights particularly difficult noisy positive pairs but is penalized by Ω to avoid unnecessary downweighting. Fig. 8 (*top right*) shows that for small overlaps (*e.g.*, 30% red box in Fig. 8 (*top left*)), our SimCLR+AMCL recovers performance, while SimCLR performs poorly. Fig. 8 (*bottom right*) shows the same trend for heavy



	30%	45%	60%	75%	90%
SimCLR	68.5	72.8	80.2	79.0	78.5
+AMCL	82.3	83.0	82.5	82.2	82.5

Color distortion strength	1/8	1/4	1/2	1
SimCLR	66.5	68.0	69.6	70.3
+AMCL	80.5	81.0	82.1	81.7

Fig. 8: (*Top left*) Overlap percentages of crops between positive pairs. (*Top right*) Evaluations of the effects of overlap percentages. (*Bottom left*) Average and variance of temperatures of positive pairs with different overlap percentages; the red curve represents the average sample-wise variance of temperatures from three heads. (*Bottom right*) Evaluation of different color distortion strengths.

color distortion. Fig. 8 (*bottom left*) shows the average (over epochs) temperature of positive-pair temperatures is high if crops overlap 30%, indicting high uncertainty. For 90% overlap, uncertainty drops.

Why not use multi-head intrinsic features consistent across different heads? This approach is handcrafted. Instead, we allow each head to specialize driven by the data, similar to multiple attention heads in a transformer. As each head is initialized differently, it captures various aspects of the data. Fig. 8 (*bottom left*) red curve shows the average sample-wise variance of temperatures from three heads. High variances indicate that the temperature of each head differs, so each head's alignment varies (global/local for high/low temperature). In experiments, a single head could not efficiently capture different aspects of the content. In contrast, multi-head captures complementary aspects of similarity between views, *e.g.*, attributes, textures, shapes, *etc.*, due to a pair-adaptive head-wise temperature (Fig. 1 (b)-(g) in the main paper), contributing to a more robust and refined similarity measure (Fig. 3 in the main paper).

Why did this method outperform SOTA? Our adaptive temperature is based on aleatoric uncertainty modeling, which adapts heads to difficult positive/negative pairs.

Why not use multiple backbones to improve feature learning? This idea has been explored in supervised learning [43]. However, training multiple backbones imposes prohibitive computational costs in SSL with no guarantee of the complementarity of such backbones.

Adaptive temperature *vs.* attention learning. The latter assigns varying weights to different components or parts of an object according to a specific design [2,8,11]. The learnable positive and negative temperatures reweigh the similarities by considering diverse image content resulting from multiple augmentations. This correction replaces the global temperature, allowing the backbone

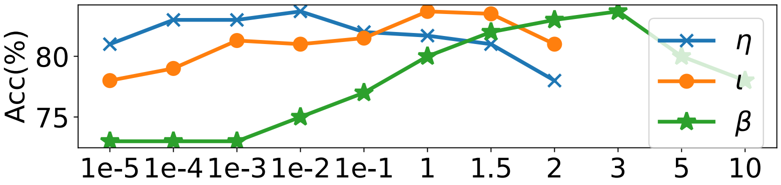


Fig. 9: Sensitivity analysis of η, ι, and β on STL-10.

and multiple projection heads to focus on capturing different aspects of image content. Moreover, pair-wise weighted similarities on 'alignment' and 'uniformity' allow various similarity relations to contribute differently to contrastive learning, similar to an attention learning mechanism.

Sensitivity analysis of η, ι, and β. We use Hyperopt package for hyperparameter optimization, running a total of 25 iterations. The search spaces for η, ι, and β are $[1e{-}5, 2]$, $[1e{-}5, 2]$, and $[1e{-}5, 10]$, respectively, as mentioned in the main paper. Fig. 9 shows the sensitivity analysis of η, ι, and β on the STL-10 dataset.

C Dataset details

We choose popular datasets that are widely used in evaluating the SSL models.

CIFAR-10 [33] consists of 60,000 32×32 colour images divided into 10 classes, each containing 6,000 images. The dataset is split into 50,000 training images and 10,000 test images.

CIFAR-100 is similar to CIFAR-10 but comprises 100 classes, each with 600 images. There are 500 training images and 100 testing images per class. The 100 classes in CIFAR-100 [33] are grouped into 20 superclasses. Each image is labeled with both a 'fine' label (indicating its specific class) and a 'coarse' label (indicating its superclass).

STL-10 [12] is similarly to CIFAR-10 and includes images from 10 classes: airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck. This dataset is relatively large and features a higher resolution (96×96 pixels) compared to CIFAR10. It also provides a substantial set of $100,000$ unlabeled images that are similar to the training images but are sampled from a wider range of animals and vehicles. This makes the dataset ideal for showcasing the benefits of self-supervised learning.

Tiny-ImageNet [35] contains 100,000 images of 200 classes (500 for each class) downsized to 64×64 colored images. Each class has 500 training images, 50 validation images, and 50 test images.

ImageNet [13] (a.k.a. **ImageNet-1K**) contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection. The publicly released dataset contains a set of manually annotated training images.

D Impact statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.