Supplementary Material:

Cost-Sensitive Learning of Deep Feature Representations from Imbalanced Data

S. H. Khan, M. Hayat, M. Bennamoun, F. Sohel and R. Togneri

APPENDIX A
PROOFS REGARDING COST MATRIX ξ'

Lemma A.1. *Offsetting the columns of the cost matrix ξ' by any constant 'c' does not affect the associated classification risk $\mathcal{R}$.*

Proof: From Eq. 1, we have:

$$\sum_q \xi'_{p^*,q} P(q|\mathbf{x}) \le \sum_q \xi'_{p,q} P(q|\mathbf{x}) \quad \forall p \neq p*$$

which gives the following relation:

$$P(p^*|\mathbf{x}) \left(\xi'_{p^*,p^*} - \xi'_{p,p^*}\right) \le \sum_{q \neq p^*} P(q|\mathbf{x}) \left(\xi'_{p,q} - \xi'_{p^*,q}\right), \quad \forall p \neq p*$$

As indicated in Sec. 3.1, the above expression holds for all $p \neq p*$. For a total number of N classes and an optimal prediction p^*, there are $N-1$ of the above relations. By adding up the left and the right hand sides of these $N-1$ relations we get:

$$P(p^*|\mathbf{x}) \left((N-1)\xi'_{p^*,p^*} - \sum_{p \neq p^*} \xi'_{p,p^*}\right) \le \sum_{q \neq p^*} P(q|\mathbf{x}) \left(\sum_{p \neq p^*} \xi'_{p,q} - (N-1)\xi'_{p^*,q}\right),$$

This can be simplified to:

$$\mathbf{P_x} \begin{bmatrix} \sum_i \xi'_{i,1} - N\xi'_{p^*,1} \\ \vdots \\ \sum_i \xi'_{i,N} - N\xi'_{p^*,N} \end{bmatrix} \ge 0,$$

where, $\mathbf{P_x} = [P(1|\mathbf{x}), \ldots, P(N|\mathbf{x})]$. Note that the posterior probabilities $\mathbf{P_x}$ are positive ($\sum_i P(i|\mathbf{x}) = 1$ and $P(i|\mathbf{x}) > 0$). It can be seen from the above equation that the addition of any constant c, does not affect the overall relation, i.e., for any column j,

$$\sum_i (\xi'_{i,j} + c) - N(\xi'_{p^*,j} + c) = \sum_i \xi'_{i,j} - N\xi'_{p^*,j}$$

Therefore, the columns of the cost matrix can be shifted by a constant c without any effect on the associated risk. ■

Lemma A.2. *The cost of the true class should be less than the mean cost of all misclassification.*

Proof: Since, $\mathbf{P_x}$ can take any distribution of values, we end up with the following constraint:

$$\sum_i \xi'_{i,j} - N\xi'_{p^*,j} \ge 0, \quad j \in [1, N].$$

S. H. Khan is with Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO) and College of Engineering & Computer Science, Australian National University, Canberra, ACT 0200, Australia. E-mail: salman.khan @data61.csiro.au

M. Bennamoun is with the School of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia.
E-mail: mohammed.bennamoun @uwa.edu.au

M. Hayat is with Human-Centered Technology Research Centre, University of Canberra, Bruce, Australia.
Email: munawar.hayat@canberra.edu.au

R. Togneri is with the School of Electrical, Electronic and Computer Engineering, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia.
E-mail: roberto.togneri@uwa.edu.au

F. Sohel is with the School of Engineering and Information Technology, Murdoch University, 90 South St, Murdoch WA 6150, Australia.
E-mail: f.sohel@murdoch.edu.au

For a correct prediction p^*, $P(p^*|\mathbf{x}) > P(p|\mathbf{x}), \forall p \neq p^*$. Which implies that:

$$\xi'_{p^*,p^*} \leq \frac{1}{N} \sum_i \xi'_{i,p^*}.$$

It can be seen that the cost insensitive matrix (when diag$(\xi') = 0$ and $\xi'_{i,j} = 1, \forall j \neq i$) satisfies this relation and provides the upper bound. ■

Lemma A.3. *The cost matrix ξ for a cost-insensitive loss function is an all-ones matrix, $\mathbf{1}^{p\times p}$, rather than a $\mathbf{1} - \mathbf{I}$ matrix, as in the case of the traditionally used cost matrix ξ'.*

Proof: With all costs equal to the multiplicative identity i.e., $\xi_{p,q} = 1$, the CNN activations will remain unchanged. Therefore, all decisions have a uniform cost of 1 and the classifier is cost-insensitive. ■

Lemma A.4. *All costs in ξ are positive, i.e., $\xi \succ 0$.*

Proof: We adopt a proof by contradiction. Let us suppose that $\xi_{p,q} = 0$. During training in this case, the corresponding score for class q $(s_{p,q})$ will always be zero for all samples belonging to class p. As a result, the output activation (y_q) and the back-propagated error will be independent of the weight parameters of the network, which proves the Lemma. ■

Lemma A.5. *The cost matrix ξ is defined such that all of its elements in are within the range $(0, 1]$, i.e., $\xi_{p,q} \in (0, 1]$.*

Proof: Based on Lemmas A.3 and A.4, it is trivial that the costs are with-in the range $(0, 1]$. ■

Lemma A.6. *Offsetting the columns of the cost matrix ξ can lead to an equally probable guess point.*

Proof: Let us consider the case of a cost-insensitive loss function. In this case, $\xi = \mathbf{1}$ (from Lemma A.3). Offsetting all of its columns by a constant $c = 1$ will lead to $\xi = \mathbf{0}$. For $\xi = \mathbf{0}$, the CNN outputs will be zero for any $\mathbf{o}^{(i)} \in \mathbb{R}^N$. Therefore, the classifier will make a random guess for classification. ■