DHN-NCE vectorisation

$$\mathcal{L} = \mathcal{L}^{v\to t} + \mathcal{L}^{t\to v}$$ with hyper-parameters $\tau, \beta_1, \beta_2 \in \mathbb{R}$.

$$\mathcal{L}^{v\to t} = -\sum_{i=1}^{B} \frac{I_{pi}^T T_{pi}}{\tau} + \sum_{i=1}^{B} \log \sum_{j\neq i} \exp\left(\frac{I_{pi}^T T_{pj}}{\tau}\right) W^{v\to t}_{I_{pi} T_{pj}}$$

Since the weight matrix W's leading diagonal $(i=j)$ is not included in the 2nd term, we can manually set it to -1 to merge the 2 terms:

$$\mathcal{L}^{v\to t} = \sum_{i=1}^{B} \log \sum_{j=1}^{B} \exp\left(\frac{I_{pi}^T T_{pj}}{\tau}\right) W^{v\to t}_{I_{pi} T_{pj}}$$

where $W_{ii} = -1 \ \forall i$. We set this manually in the final setting.

Set $S_{ij} = \frac{I_{pi}^T T_{pj}}{\tau}$ where $I \in \mathbb{R}^{B\times n}$, $T \in \mathbb{R}^{B\times n}$ are the batch matrices of each image and text vector.

So $S = \frac{1}{\tau} I T^T \in \mathbb{R}^{B\times B}$ is a similarity matrix.

We normalise I and T to prevent gradient blow-up.

Let $A = \exp(S)$

So $$\mathcal{L}^{v\to t} = \sum_{i=1}^{B} \log \sum_{j=1}^{B} AW$$ with $\mathrm{diag}(W) = -1$.

Note that for $\mathcal{L}^{t\to v}$ we calculate it by $T_{pi}^T I_{pj}$, ~~which transposes~~ and $T_{pi}^T I_{pj} = (I_{pi}^T T_{pj})^T$, so we use S^T in our calculations instead of S to calculate $\mathcal{L}^{t\to v}$.

$$W = (B-1) \frac{\exp\left(\frac{\beta_1 I_{pi}^T T_{pj}}{\tau}\right)}{\sum_{k\neq i} \exp\left(\frac{\beta_1 I_{pi}^T T_{pk}}{\tau}\right)} = (B-1) \frac{A_{ij}^{\beta_1}}{\sum_{k\neq i} A_{ij}^{\beta_1}} = (B-1) \frac{e^{\beta_1 S_{ij}}}{\sum_{k\neq i} e^{\beta_1 S_{ik}}}$$

We ~~can~~ can define $X = \beta_1 S$ element-wise and set $X_{ii} = -\infty \ \forall i$.

Then $W = (B-1)\, \mathrm{softmax}(X, \dim=-1)$.

This would be $\dim = 0$ for $W^{t\to v}$, but transposing $S \Rightarrow$ transposing X does the same

So
1. Calculate $S = \frac{1}{\tau} I T^T$
2. Calculate $A = \exp(S)$
3. Calculate $X_1 = \beta_1 S$, set $X_{ii} = -\infty$
4. Calculate $W = (B-1) \cdot \text{softmax}(X, \text{dim}=-1)$
5. Set $W_{ii} = -1$
6. $\mathcal{L}^{V \to t} = \text{sum}(\log(\text{sum}(A \odot W)))$

Then do the same for $\mathcal{L}^{t \to v}$ by transposing S and using β_2.