# DHN-NCE Loss

We are given the loss $\mathcal{L} = \mathcal{L}^{v\to t} + \mathcal{L}^{t\to v}$, with hyperparameters $\tau, \beta_1, \beta_2$, and batch size $B$.

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

$$W^{v\to t}_{I_{pi}T_j} = (B-1)\cdot \frac{\exp\left(\frac{\beta_1 I_{pi}{}^{T} T_{pj}}{\tau}\right)}{\sum_{k\neq i} \exp\left(\frac{\beta_1 I_{pi}{}^{T} T_{pk}}{\tau}\right)}$$

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

$$W^{t\to v}_{T_{pi}I_{pj}} = (B-1)\cdot \frac{\exp\left(\frac{\beta_2 T_{pi}{}^{T} I_{pj}}{\tau}\right)}{\sum_{k\neq i} \exp\left(\frac{\beta_2 T_{pi}{}^{T} I_{pj}}{\tau}\right)}$$

We aim to vectorise these operations which leverage a GPU.
We start with matrices $I, T \in \mathbb{R}^{B\times n}$ which are the matrices of text and image embeddings calculated from the model. These represent 1 batch.
We aim to find $f(I, T, \tau, \beta_1, \beta_2)$ which is the loss of matrices $I, V$.

We will leverage the summations that skip the leading diagonal.

Fist $I, T \in \mathbb{R}^{B\times n}$.

Define similarity matrix $S = I^{T} T$.
and $A = \exp\left(\frac{1}{\tau} S\right)$

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

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

As mentioned, the summation ignores the case when $j = i$, which is the leading diagonal of matrix A. So we can add the $A_{ii}$ leading diagonal in, but set its weight to -1 to represent the coefficients.

$$\text{So } \mathcal{L}^{v\to t} = \sum_{i=1}^{B} \log \sum_{j=1}^{B} A_{ij} W_{ij}^{v\to t}$$

Then we have $W_{ij}^{v\to t} = (B-1)\, \frac{\exp\left(\frac{\beta_1 I_{pi}^{T} T_{pj}}{\tau}\right)}{\sum_{k\neq i} \exp\left(\frac{\beta_1 I_{pi}^{T} T_{pj}}{\tau}\right)} \quad \in \mathbb{R}^{B\times B}$

is equivalent to

The fraction ~~would be~~ a softmax function with the leading diagonal equal to 0. ~~And if $i=j$, we will replace~~ And $W_{ii}^{v\to t} = -1 \quad \forall i \in \{1,\dots,B\}$ anyway. So, in calculating $W_{S_{pi}T_{pj}}^{v\to b}$, we set ~~diag(W) = 0~~ the diagonal to 0, apply softmax and then ~~multiply by 1~~. subtract 1 from the diagonal.

$$W_{ij}^{v\to t} = (B-1) \cdot \text{softmax}\left(A_{ij}^{\beta_1} - I_B A_{ij}^{\beta_1}, \text{dim}=-1\right) - I_B \quad \in \mathbb{R}^{B\times B}$$

$A_{ij}^{\beta_1} - I_B A_{ij}^{\beta_1}$ applies the $\beta_1$ scaling element-wise and then replaces the leading diagonal with zeroes.

$$\text{So } \mathcal{L}^{v\to t} = \sum_{i=1}^{B} \log \sum_{j=1}^{B} A_{ij} \otimes \left\{(B-1)\,\text{softmax}\left(A_{ij}^{\beta_1} - I_B A_{ij}^{\beta_1}, \text{dim}=-1\right) - I_B\right\}.$$

We use the Hadamaard product since we apply a linear combination to each image, resulting in B total computations. A matrix multiplication will require $B^2$ computations, which is inefficient.

Since there is a log term between the sums, we cannot turn this into a matrix multiplication as it is non-linear.

~~As log is one-to-one and strictly increasing removing the log~~
~~So $\mathcal{L}^{v\to t}$ = sum (log ( sum ( A ⊗ (B-1) softmax ( $A_{ij}^{\beta_1} - I_B A$~~

So $\mathcal{L}^{v \to t} = \text{sum}\left(\log\left(\text{sum}\left(A \otimes \{(B-1)\,\text{softmax}\left(A^{\beta_1} - I_B A^{\beta_1}, dim=-1\right) - I_B\}\right)\right)\right)$

By inspection if we say $\mathcal{L}^{v \to t} = g(A)$ (and $A = h(I, T)$),
$\mathcal{L}^{t \to v} = g(A^T)$. and vice versa.

So $\mathcal{L} = \text{sum}\left(\log\left(\text{sum}\left[A \otimes \{(B-1)\,\text{softmax}\left(A^{\beta_1} - I_B A^{\beta_1}, dim=-1\right) - I_B\}\right]\right)\right)$
$+ \text{sum}\left(\log\left(\text{sum}\left[A^T \otimes \{(B-1)\,\text{softmax}\left((A^{\beta_2})^T - I_B A^{\beta_2 T}, dim=-1\right) - I_B\}\right]\right)\right)$

where $A = \exp\left(\frac{I^T T}{\tau}\right)$ and $A^{\beta_1}, A^{\beta_2}$ are applied element-wise.

For reference, $A \otimes W \in \mathbb{R}^{B \times B}$
$\text{sum}(A \otimes W) \in \mathbb{R}^{B \times 1}$ as it is applied on 1 axis
$\log(\text{sum}(A \otimes W)) \in \mathbb{R}^{B \times 1}$
$\text{sum}(\log(\text{sum}(A \otimes W))) \in \mathbb{R}^{1 \times 1}$