A. Comparing ΔI and I

We propose $\Delta I(L, C^{s_1}, ..., C^{s_K})$ to approximate the Interaction Information between K context features and the network loss L, $I(L, C^{s_1}, ..., C^{s_K})$. The computational complexity of computing $\Delta I(L, C^{s_1}, ..., C^{s_K})$ grows linearly with K, as compared to the computational complexity of computing $I(L, C^{s_1}, ..., C^{s_K})$ which grows combinatorially with K. We investigate the difference between $I(L, C^{s_1}, ..., C^{s_K})$ and $\Delta I(L, C^{s_1}, ..., C^{s_K})$. To simplify the notation, we denote C^{s_1} as C^1 and C^{s_2} as C^2. It is trivial to compute the Mutual Information between the context features and L and select C^1 to be the feature most informative about the loss. We assume C^1 has been selected and we compare $I(L, C^1, C^2)$ and $\Delta I(L, C^1, C^2)$.

$$I(L, C^1, C^2) = I(L, C^2) - I(L, C^2|C^1) \tag{11}$$

$$\Delta I(L, C^1, C^2) = I(L, C^2) - I(C^1, C^2) \tag{12}$$

The difference between $I(L, C^1, C^2)$ and $\Delta I(L, C^1, C^2)$ is:

$$I(L, C^1, C^2) - \Delta I(L, C^1, C^2) = I(C^1, C^2) - I(L, C^2|C^1) \tag{13}$$

As we would like the context features in $\mathbf{C}^{S_K}$ to be roughly independent, let us assume that C^1 is not informative of C^2, i.e., $I(C^1, C^2) = 0$.

$$I(L, C^1, C^2) - \Delta I(L, C^1, C^2) = -I(L, C^2|C^1) \tag{14}$$

The reader is reminded that the conditional mutual information is computed as:

$$I(L, C^2|C^1) = \sum_{\ell \in L} \sum_{c_1 \in C_1} \sum_{c_2 \in C_2} p(\ell, c_1, c_2) \times log\left(\frac{p(c_1)p(\ell, c_1, c_2)}{p(\ell, c_1)p(c_1, c_2)}\right) \tag{15}$$

For simplicity, let us consider the point wise conditional mutual information at ℓ, c_1, and c_2:

$$log\left(\frac{p(c_1)p(\ell, c_1, c_2)}{p(\ell, c_1)p(c_1, c_2)}\right) \tag{16}$$

Recall, it was assumed that C^1 and C^2 are independent, thus $p(c_1, c_2) = p(c_1)p(c_2)$. The joint probability $p(\ell, c_1, c_2)$ can also be factored as $\frac{1}{Z}\psi(\ell, c_1)\psi(\ell, c_2)$.

$$= log\left(\frac{p(c_1)\psi(\ell, c_1)\psi(\ell, c_2)}{Zp(\ell, c_1)p(c_1)p(c_2)}\right) \tag{17}$$

$$= log\left(\frac{\psi(\ell, c_1)\psi(\ell, c_2)}{Zp(\ell, c_1)p(c_2)}\right) \tag{18}$$

Note $\psi(\ell, c_1) \propto p(\ell, c_1)$ and $\psi(\ell, c_2) \propto p(\ell, c_2)$. Thus, the difference between the proposed ΔI and the Interaction Information is proportional to

$$\propto log\left(p(\ell|c_2)\right) \tag{19}$$

If we consider only combinations of ℓ and c_2 that exist in the test set, $p(\ell|c_2) > 0$. As the new context feature becomes more informative, $p(\ell|c_2) \to 1$ and the difference $log\left(p(\ell|c_2)\right) \to 0$. This demonstrates that, if the context features are informative about the loss, ΔI is a good approximation of the Interaction Information.