**Review：Unsupervised Learning of Visual Features by Contrasting Cluster Assignments**

SwAV proposes a new self-supervised learning paradigm to learn feature representation by comparing clustering assignments of different views, avoiding direct feature comparison.

**Main idea:**

* comparing clustering assignments of different views instead of directly comparing sample features
* Online clustering mechanism is introduced to dynamically learn prototype vectors.
* Sinkhorn-Knopp algorithm is used to optimize clustering assignment.
* Multi-crop strategy is proposed to significantly improve performance.

**The training process of SwAV consists of the following steps:**

1. Data augmentation: Generate two different augmented views x\_t, x\_s for each image
2. Feature extraction: Get features z\_t = f(x\_t), z\_s = f(x\_s) through encoder f
3. Prototype mapping: Match features with a set of prototype vectors {c1,...,cK} to get scores
4. Encoding calculation:

* Use Sinkhorn-Knopp algorithm to calculate soft clustering assignment to get encoding q\_t, q\_s
* Use softmax to calculate probability distribution to get p\_t, p\_s

1. Cross prediction: Minimize cross prediction loss L(z\_t, z\_s) = l(z\_t, q\_s) + l(z\_s, q\_t) Where l(z,q) measures the degree of fit of feature z to the prediction of encoding q

**Formula derivation and description:**

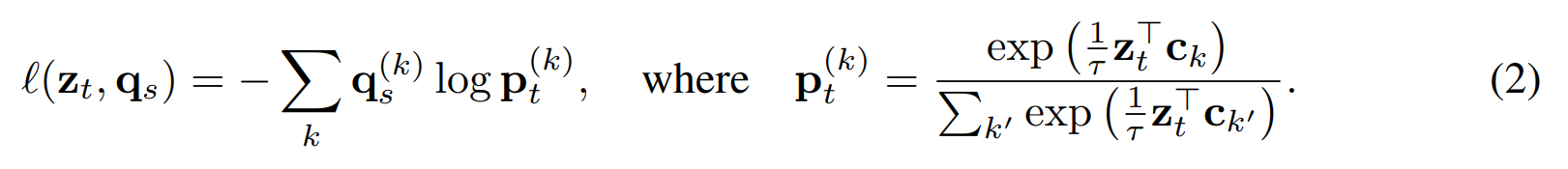
L(zt, zs) = ℓ(zt, qs) + ℓ(zs, qt)



The core loss function in SwAV.

zt and zs: the feature representations of two different data-augmented views of the same image

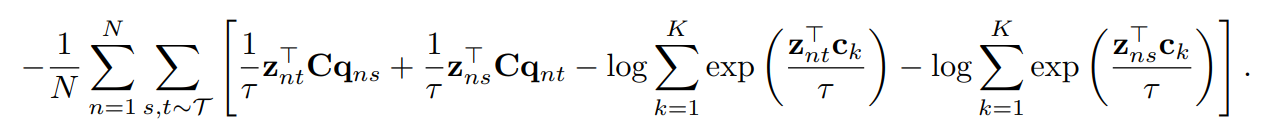
qt and qs: the "encodings" (cluster assignment probabilities) corresponding to these two features.



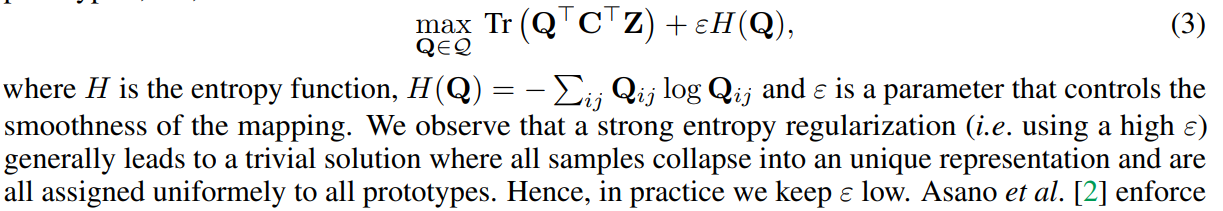
q\_s^(k): the soft clustering assignment obtained by the Sinkhorn-Knopp algorithm

p\_t^(k): the probability distribution of the similarity between feature z\_t and prototype c\_k after softmax

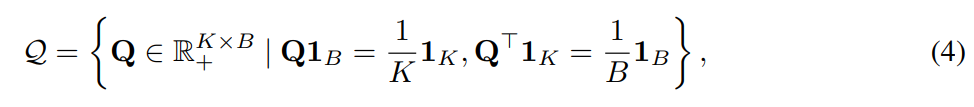
τ: a temperature parameter used to adjust the smoothness of the distribution



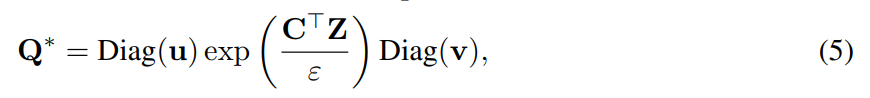
Average the entire batch of N samples and all possible augmentation pairs (s, t)



The term Tr(Q^T C^T Z) maximizes the similarity between feature Z and prototype C.



where 1K denotes the vector of ones in dimension K. These constraints enforce that on average each prototype is selected at least B K times in the batch.



Q\* is the optimal solution to the optimization problem (3), that is, the optimal soft clustering assignment matrix, which is a K×B matrix that represents the soft assignment probability of B samples to K prototypes. u and v are renormalized vectors with dimensions K and B respectively

u = ones(K)/K

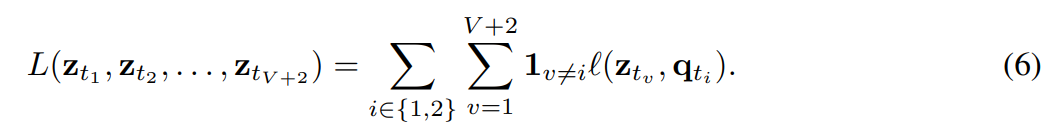
v = ones(B)/B

exp(C^T Z/ε) calculates the similarity between the feature and the prototype

Diag(u) and Diag(v) ensure that the equilibrium constraint is met

Use the Sinkhorn-Knopp algorithm to iteratively calculate the renormalized vectors u and v

Only 3 iterations are needed to get good results



z\_t1, z\_t2 are two full-resolution views

z\_t3 to z\_t(V+2) are V low-resolution views

1\_{v≠i} means not comparing with itself

The above formula is equal to:

# Comparison between full resolution views+Comparison between low resolution view and full resolution view

l(z\_t1, q\_t2) + l(z\_t2, q\_t1) +∑\_{v=3}^{V+2} [l(z\_tv, q\_t1) + l(z\_tv, q\_t2)]