Relational Proxies: Emergent Relationships as Fine-Grained Discriminators

Appendix

Abhra Chaudhuri[1] **Massimiliano Mancini**[2] **Zeynep Akata**[2,3,4] **Anjan Dutta**[5]*
[1] University of Exeter [2] University of Tübingen [3] MPI for Informatics
[4] MPI for Intelligent Systems [5] University of Surrey

1 Properties of the Relationship Modelling Function

Intuitive Analogy: The problem of local-to-global relation computation can be viewed as a bit-string-to-integer matching problem. Consider 3 bits, say b_1, b_2 and b_3, corresponding to 3 local views. Let the global view be represented by an integer that can be encoded with 3 bits, say with a value of g = 6, for this example. The problem then is to find the association of the integer 6 with its corresponding binary representation of 110. This association represents the cross-view relationship.

The first step towards solving this problem is to *enumerate* all the possible ways in which the local views can combine (to produce any global view, not specifically g). The set of all such combinations will be given by $S = \{000, 001, 010, ..., 110, 111\}$. The bit values encode the presence or absence of a particular view in the cross-view relationship. So, no matter what order we observe b_1, b_2 and b_3 in, we must output the same set S, as it is required to be an exhaustive enumeration. This is exactly what the property of permutation invariance achieves. Once we have S, the next step is to *find* the mapping $S, g \mapsto 110$, i.e, the correct binary encoding for the integer $g = 6$, which is accomplished by the property of view-unification.

Purpose: As illustrated through the above analogy, one can view the local-to-global relationship modelling function as an enumerative search algorithm - given a set of local views, it first *enumerates* all possible ways in which they can combine to form a meaningful global view. Given that enumeration, it then *finds* the target solution by learning to identify the correct combination that matches with the global-view representation. Thus, the enumerate operation needs to be permutation invariant, as it has to consider all possible combinations of the inputs, and the find operation needs to be a view-unifier by construction.

Motivation: Behind our specific design choice was the motivation to keep the enumerate and find steps separate. This allows the model to have dedicated representation spaces for the two distinct sub-tasks, which in turn facilitates better convergence.

2 Proofs of Additional Identities

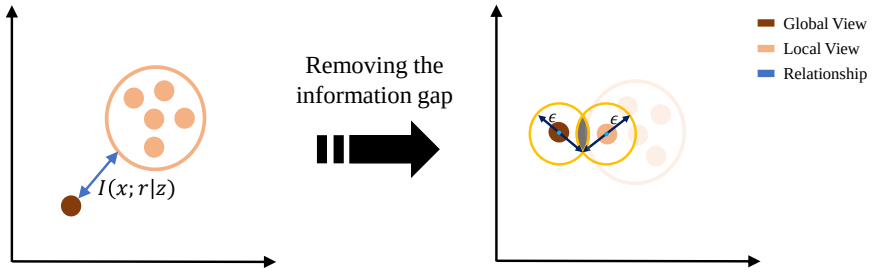
Identity 1. *Given a relation-agnostic representation* $\mathbf{z}$ *of* $\mathbf{x}$*, the only uncertainty that remains about the label information* $\mathbf{y}$ *can be quantified as the cross-view relational information* $\mathbf{r}$*, i.e.,* $I(\mathbf{x}; \mathbf{y}|\mathbf{z}) = I(\mathbf{x}; \mathbf{r})$.

Proof. Using the chain rule for mutual information [3], we can factorize the label information $\mathbf{y}$ contained in $\mathbf{x}$, *i.e.*, $I(\mathbf{x}; \mathbf{y})$ as:

$$I(\mathbf{x}; \mathbf{y}) = I(\mathbf{x}; \mathbf{y}|\mathbf{z}) + I(\mathbf{x}; \mathbf{z}) \tag{1}$$

As evidenced by recent literature [4, 5, 1, 2], the label information in $\mathbf{x}$ can be expressed exclusively as a function of its global ($\mathbf{g}$) and local ($\mathbf{l}_i$) views. Thus, in quantitative terms, the label information

*A. Chaudhuri is with the Department of Computer Science at the University of Exeter. M. Mancini and Z. Akata are with the Cluster of Excellence Machine Learning at the University of Tübingen. A. Dutta is with the Institute for People-Centred AI at the University of Surrey.
36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) (b)

Figure 1: (a) A relation-agnostic representation space. (b) The ϵ-neighborhoods of the global and local views begin colliding as the information gap is reduced.

$\mathbf{y}$ in $\mathbf{x}$ can also be factorized into *relation-agnostic* and *relation-aware* components as follows:

$$I(\mathbf{x};\mathbf{y}) = \underbrace{I(\mathbf{x};\mathbf{g}) + \sum_{\mathbf{l}\in\mathbb{L}} I(\mathbf{x};\mathbf{l})}_{\text{relation-agnostic}} + \underbrace{I(\mathbf{x};\mathbf{r})}_{\text{relation-aware}} \tag{2}$$

The relation-aware representation $\mathbf{r}$ is, unlike relation-agnostic representations, obtained explicitly based on the cross-view relationship. However, since f computes $\mathbf{z}$ without considering any relational information, it only models the relation-agnostic component of Equation (2). Thus,

$$I(\mathbf{x};\mathbf{z}) = I(\mathbf{x};\mathbf{g}) + \sum_{\mathbf{l}\in\mathbb{L}} I(\mathbf{x};\mathbf{l}) \tag{3}$$

Substituting the *relation-agnostic* component of Equation (2) with the L.H.S. of Equation (3), and comparing it with Equation (1), we get:

$$I(\mathbf{x};\mathbf{y}|\mathbf{z}) = I(\mathbf{x};\mathbf{r}) \tag{4}$$

□

3 Geometric Relation Agnosticity

Definition 4 is based on the fact that the information gap (derived in Proposition 1) between the global and the local views has the effect that the two view families would be mapped to distinct locations in the representation space, and the separation between them would be proportional to the information gap, *i.e.*, $I(\mathbf{x};\mathbf{r}|\mathbf{z})$. Definition 4 also mentions that relation-agnostic embeddings of the local and the global views must thus be well separated, *i.e.*, the ϵ-neighborhood of the global embedding $n_\epsilon(\mathbf{z}_g)$ must not intersect with those of the local embeddings $n_\epsilon(\mathbf{z}_l)$. In other words, the global embedding must be sufficiently far apart from each of the local embeddings.

Figure 1 depicts the geometric effect of removing the information gap from a relation-agnostic representation space. As proven in Lemma 3, if the information gap is reduced using the same encoder f that was used to obtain $\mathbf{z_g}$ and $\mathbf{z}_l$, the model starts mapping the global and the local views to identical regions in the representation space. This could potentially lead to the requirement of k-distinguishability to not be satisfied, as the unique information pertaining to at least one of the local views is lost upon merger with the global view (and vice-versa). It is thus a requirement for a sufficient learner to preserve the relation-agnosticity in the representation space of f.

4 Relation-Agnosticity of Relational Proxies

The representations $\mathbf{z}_g$ and $\mathbf{z}_{l_i}$ are computed in a relation-agnostic manner and no explicit operation is performed to reduce the domain gap between the global and the set of local views. This natural domain gap thus manifests in the representation space of $\mathbf{z}$ as its relation-agnostic nature.

Figure 2 diagrammatically illustrates this idea. Given an entangled representation space where the classes are not entirely separable (left), the encoder f has two choices to map the local and global views of the corresponding datapoints to completely separable proxy neighborhoods. It could either:

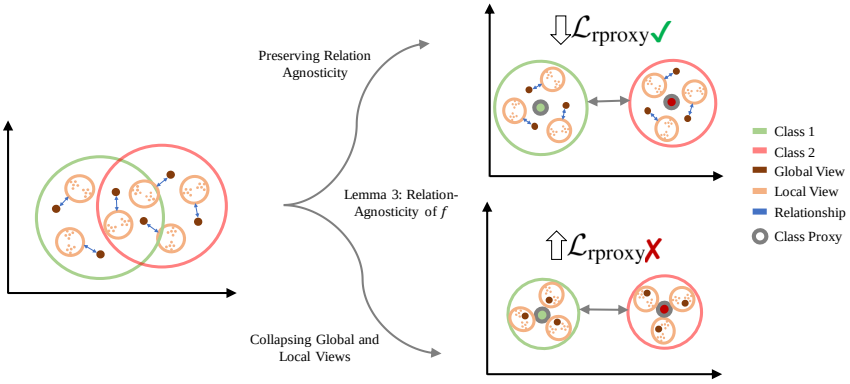


Figure 2: Embeddings of datapoints from two classes obtained from f before training (left). As f is trained with an end objective of minimizing $\mathcal{L}_{\text{rporxy}}$, it has two potential choices (right). However, as proven in Lemma 3, the relation-agnostic nature of f prevents the collapse of the global and local embeddings even when they share the same set of class proxies.

1. Preserve the relation-agnosticity by maintaining the information gap (equal to the cross-view relational information) even within the proxy neighborhood (top right), or
2. Collapse the local and global representations in the process of alignment (bottom right) by mapping them to ϵ-neighborhoods of each other.

However, since the end objective of our model is to minimize $\mathcal{L}_{\text{rproxy}}$, which is cross-entropic in nature, we prove via Lemma 3 that f cannot collapse the local and global representations, as that would lead to an increase in the downstream cross-entropy loss. f would thus choose to preserve the relational gap in the representation space while mapping them to the neighborhood of their corresponding proxy.

5 Visual Representations of Cross-View Local Relationships

Figure 3 depicts examples of graphs depicting cross-view local relationships. It can be seen that images that provide a diverse set of local views, and thus, a larger space of possible cross-view relationships are the ones that get classified correctly with full certainty. However, as the number of unique local views get limited (possibly due to occlusion or an incomplete photographing of the object), it reduces the amount of relational information that can be mined. Under situations when even the individual local-views are largely shared between classes, there remains no discriminative premise (neither local/global, nor relational) for telling their instances (with limited depiction of local views) apart. It is under such circumstances that the classifier gets confused.

Example: For instance, in the example from the CUB dataset (the top row in Figure 3), the images of the Acadian Flycatcher and Bank Swallow depict sufficient numbers of local views like the head, tail, belly and wings, which provide a large space of potential cross-view relationships that favor classification outcome. On the other hand, the images of the Black-footed Albatross and Laysan Albatross only depict the head and the neck, thus limiting the number of computable relationships that can act as discriminators. Moreover, the head and the neck look largely similar between the two categories, thereby leading to cross-category confusion causing a subsequent misclassification. However, we believe that such a situation can be addressed by learning different distributional priors over the set of local views, which we plan to take up as future work.

References

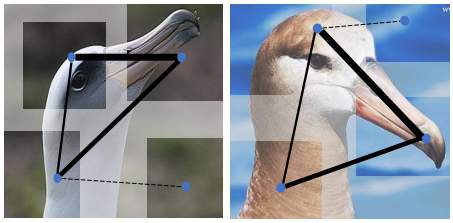
[1] Ardhendu Behera, Zachary Wharton, and Asish Bera. Context-aware Attentional Pooling (CAP) for Fine-grained Visual Classification. In *AAAI*, 2021.

[2] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised Part Discovery from Contrastive Reconstruction. In *NeurIPS*, 2021.

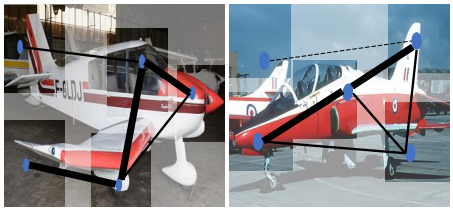
[3] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning Robust Representations via Multi-View Information Bottleneck. In *ICLR*, 2020.

[4] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry Davis, Jun Li, Jian Yang, and Ser Nam Lim. Cross-x learning for fine-grained visual categorization. In *ICCV*, 2019.

Correct Classifications: Acadian Flycatcher (left) and Bank Swallow (right).

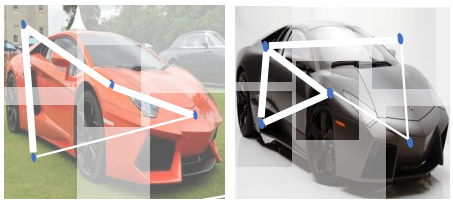


Misclassifications: Black-footed Albatross (left) and Laysan Albatross (right) confused with each other.

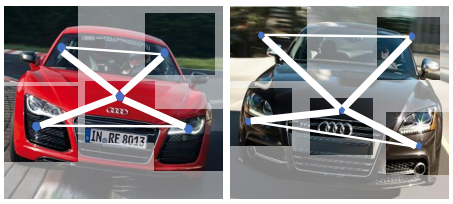


Correct Classifications: DR-400 (left) and Hawk T1 (right).

Misclassifications: Boeing-727 (left) and Falcon-900 (right) confused with each other.



Correct Classifications: Lamborghini Aventador (left) and Lamborghini Reventón (right).



Misclassifications: Audi R8 (left) and Audi TT (right) confused with each other.

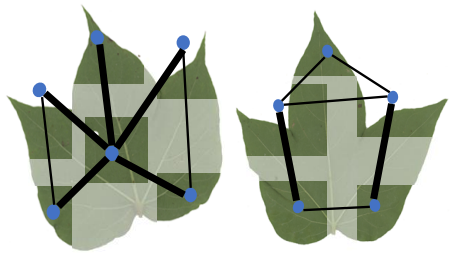


Figure 3: Visual representations of cross-view relationships along with qualitative classification results on (in order from top) CUB, FGVC Aircraft, Stanford Cars and Cotton Cultivar datasets. The pairs on the left correspond to correct classifications made by our model, while the ones on the right are misclassifications occurring out of cross-category confusions.

[5] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization. In *MMM*, 2021.