Weakly Supervised Learning via Relational Comparisons

Sina Mollaei
Stanford University
mollaei@stanford.edu

Junha Lee
Stanford University
junhalee@stanford.edu

Abstract

Obtaining labels for fully-supervised learning is often a cost and labor-intensive task. In many settings, it is easier to obtain a pairwise relationship (i.e., the label for A is greater than the label of B). We investigate the efficacy of training in this weakly-supervised setting. Using the MNIST handwriting dataset as a toy example, we examine 3 models. The first is a CNN classifier for the 10 class labels, used as the baseline metric for comparison. The second is a Siamese CNN model that is trained on relational data to output relationship classifications (greater than, less than). We use a transfer learning technique to then convert this Siamese model to a classifier, and train it for few-shot learning. The final model is a CNN classifier trained on purely relational data. We observe that the Siamese few-shot learning model is the most efficient at using the pairwise-relationship data for classification. As few as 1 explicit image-label pair per class is needed to achieve 98.43% accuracy for digit classification, which is similar to the accuracy achieved with 50k datapoints in a fully supervised setting.

1. Introduction

Deep learning models have achieved remarkable performance across various computer vision tasks. However, most successful approaches rely on fully supervised learning with large labeled datasets, which require significant human annotation effort. Weakly supervised learning aims to reduce this dependence on extensive labeled data by utilizing cheaper, more abundant forms of supervision.

In this project, we explore a specific form of weak supervision: learning from relational comparisons between pairs of images. Instead of providing absolute class labels, we only specify the relative ordering between image pairs (e.g., "the digit in image A is greater than the digit in image B"). This type of supervision can be particularly valuable in domains where precise labels are difficult to obtain, but relative comparisons are easier to generate or are naturally available.

Our work addresses the following key research questions:

- Can models trained solely on relational data achieve accuracy comparable to fully supervised counterparts?
- How does the quantity of pairwise comparisons affect performance?
- How does post-training the model with a small set of labeled data change the dynamics of the accuracy and number of relational data needed?

2. Related Work

2.1. Manifold Learning and Digit Embeddings

Classical manifold learning techniques have long revealed that handwritten digits naturally form structured, low-dimensional embeddings. Methods like t-SNE [15], Isomap [14], and Locally Linear Embedding [12] consistently demonstrate that digits organize into meaningful clusters when projected to low dimensions. These unsupervised methods discover such structure post-hoc by analyzing the geometric properties of the data manifold. Notably, visualizations derived from t-SNE often reveal not just distinct clusters but also an apparent one-dimensional ordinal progression for digit datasets, where, for instance, the cluster for '0' is adjacent to '1', '1' to '2', and so on. This observed phenomenon, where an inherent ordinal structure is suggested by unsupervised embeddings, informed our hypothesis that models could be explicitly trained to capture and represent such one-dimensional ordinal relationships directly from relational inputs. Consequently, these visualization techniques also serve as a valuable tool for qualitatively assessing the degree to which a model has successfully learned the intended ordinal structure in its embedding space.

2.2. Siamese Networks and Metric Learning

Siamese networks, introduced by Bromley et al. [2], use twin networks with shared weights to learn similarity metrics between pairs of inputs. These architectures have been

successfully applied to face verification [4], one-shot learning [7], and visual tracking [1]. Our work extends this paradigm by using Siamese networks to learn from ordinal relationships rather than similarity, demonstrating that pairwise "greater than" comparisons naturally discover digit ordering without explicit numerical labels.

2.3. Learning from Pairwise Comparisons

Learning from pairwise preferences has been extensively studied in information retrieval [9] and recommendation systems [11]. In computer vision, ranking approaches have been applied to age estimation [3] and aesthetic assessment [8]. Our work differs by investigating whether pairwise comparisons alone can induce representations suitable for classification tasks. We show that models trained solely on ordinal relationships achieve competitive classification accuracy (up to 97.8%) when combined with minimal labeled data.

2.4. Transfer Learning and Few-Shot Classification

Transfer learning leverages pre-trained representations for new tasks with limited data [17]. Few-shot learning approaches, including prototypical networks [13] and MAML [6], aim to classify with minimal examples. Our "chimera" approach uniquely combines Siamese pretraining on pairwise comparisons with few-shot classification, achieving 96.64% accuracy with only 3 labeled examples per class. This demonstrates that relational pre-training creates well-separated clusters that require minimal supervision to map to semantic categories.

2.5. Weakly Supervised Learning

Zhou [18] categorizes weak supervision into three types: incomplete (missing labels), inexact (coarse labels), and inaccurate (noisy labels). Our approach falls under incomplete supervision, where we have relational information between examples but lack absolute labels. Recent work has explored learning from partial labels [5], noisy labels [10], and side information [16]. We contribute by showing that purely relational supervision can discover meaningful ordinal structure, and that scaling from 50k to 5M pairwise comparisons progressively improves both ordering quality and downstream classification performance.

3. Dataset

We use the MNIST dataset of handwritten digits to evaluate our approach. For the baseline supervised method, we use the dataset as-is, using a 5:1:1 train/val/test split. For the weakly supervised setting, we generate a synthetic relational dataset by

1. Sampling random pairs of images (x_i, x_j) from the dataset ($i \neq j$)
2. Generating relational labels z_{ij} based on the relationship of the original labels y_i, y_j. $z_ij = 1$ if $y_i > y_j$. We purposely omit examples where $y_i = y_j$ for simplicity.

We generate a validation and test set of 20k image pairs using only the images in the validation and test sets, respectively. For the training set, we generate 3 datasets of sizes 50k, 500k, and 5M to be able to test the effect of the quantity of relational data for training models.

The datasets are generated in `data/mnist/download_mnist.py`, and the dataset handling is done in `data/dataset.py`.

4. Technical Approach

4.1. Base Model

For consistency of the larger model structure and for transfer learning, we elected to use the same fundamental model structure for all tasks. We define a `BaseModel` that consists of `Conv2d` and `MaxPool` layers, and use this `BaseModel` as the embedding generation part of all models (Figure 1a). The models only differ by the final few layers (heads) attached to the base model.

4.2. Supervised Model

For the supervised model, we attach a classification head to the base model (Figure 1b). The classification head is a simple, fully-connected linear layer that takes the 128-dimensional embedding and outputs predictions for the 10 classes. The final prediction is done by running a softmax over the 10 logits.

The model is trained using a standard cross-entropy loss, with a learning rate of 10^{-3} and a weight decay of 10^{-4}. The model is trained for 100 epochs with batch size 64. The model with the best validation accuracy during training was selected as the final baseline model. This model was able to achieve a test accuracy of 99.38%.

4.3. Siamese CNN Model

Siamese model architectures can choose to have different models for the embedding of the first and second image; in this study, we choose to use the same model (shared weights) for both images (Figure 3b). Intuitively, this is reasonable as there is no inherent difference between the images based on position.

Embeddings of both images are computed through the `BaseModel`, and are then concatenated to form a 256-dimensional embedding for the pair. The new embedding is passed through a relational classifier head. The head contains two linear layers connected by the ReLU nonlinearity and dropout. The final result is a relational prediction, as a number between 0 and 1. The training objective is a binary

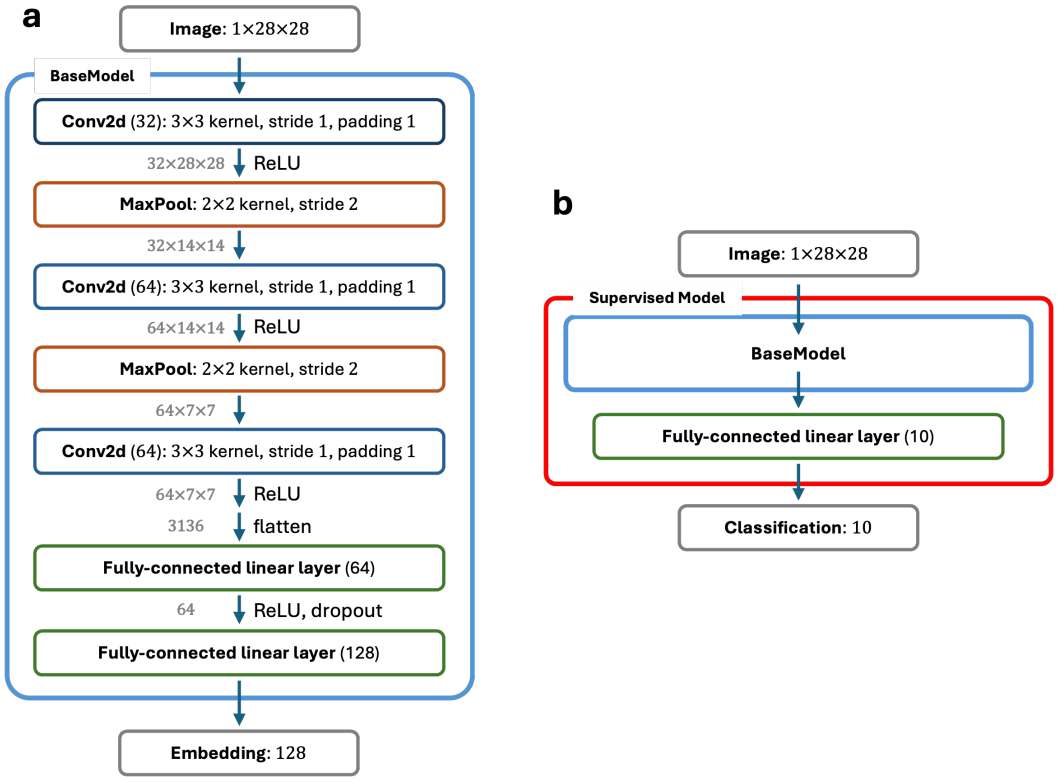

Figure 1. a) The architecture of the `BaseModel` used across all models. b) The architecture of the baseline supervised model.

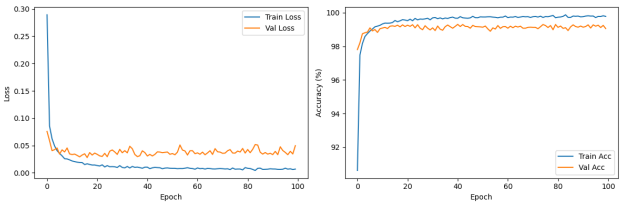

Figure 2. The loss and accuracy per epoch when training the supervised model.

Model	Test Accuracy (%)
50k	99.33
500k	99.40
5M	99.45

Table 1. Test accuracies of the Siamese models for the relational task.

cross-entropy loss between the actual label and this prediction. The Siamese models were trained with various training dataset sizes (50k / 500k / 5M image pairs). The epochs were scaled accordingly such that each setting would see the same number of datapoints (i.e., the model that uses the 50k dataset will see each datapoint 100x more times than in the 5M dataset). This was done to place a control on the number of training batches, so the effect of 'new' data could be observed.

During the training of each of the three models, we were able to achieve the following accuracies for the relational prediction task (predicting z_{ij}'s):

We then repurpose the `BaseModel` within the Siamese model as a classifier. As the model has not explicitly seen the class labels, we introduce a few-shot learning scheme: we transfer the weights of the `BaseModel` of the Siamese model into the supervised model (Figure 1b) with a randomly initialized classifier head. We then train the classifier, with the `BaseModel` weights frozen, on a small supervised training dataset. In this study, we try training on 10 datapoints (1 per label) and 30 datapoints (3 per label).

4.4. Weakly Supervised Model

The Siamese model, as will be discussed below, has shown promising results on the ability of the relational data to generate distinct segmentations in the embedding space. With this knowledge, we attempt to train a classifier fully on relational data (this is unlike the Siamese model, which had to use few-shot learning).

We achieve this by using the structure shown in Figure 3a. Each image is passed through the same `BaseModel`, which uses a fully-connected linear layer and a softmax to generate predicted probabilities for the labels of each image.

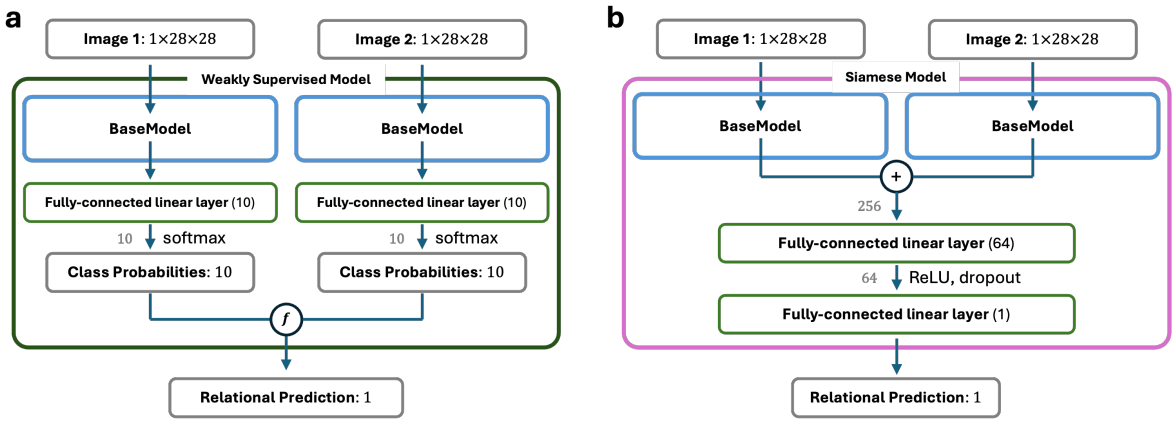

Figure 3. a) The architecture of the weakly supervised model and b) the architecture of the Siamese model.

This process is identical to the fully supervised model. To enable training on the relational data, we then use a function that uses the class probabilities of the two images to derive the probability that the first image has a greater class label than the second. This is achieved in a vectorized form by computing a cross product of the two probability vectors, then using a lower triangular mask to add the matrix elements that correspond to a probability that the first image is greater than the second. The final predicted probability is used as the predicted label.

The loss is defined as a binary cross-entropy loss, and the model is trained on 3 distinct datasets of varying sizes, as discussed for the Siamese model. The final test accuracies for the relation task (predicting z_{ij}) are listed below. We observe that the weakly supervised model performs worse on the relational task compared to the Siamese model; this is expected as we introduced a rigid inductive bias into the model (classification into 10 classes before prediction of relation).

Model	Test Accuracy (%)
50k	95.90
500k	98.23
5M	96.89

Table 2. Test accuracies for the weakly supervised model for the relational task.

By the design of this architecture, predicting actual classification labels is trivial. Instead of comparing the class probabilities of two images, we directly use the predicted probabilities as the class label predictions.

5. Results and Discussion

5.1. Supervised Model

As is known, the supervised model (baseline model) performs very well. The confusion matrix in Figure 4 shows

Model	Test Accuracy (%)
Supervised	99.23
Siamese (50k), 1-shot	97.80
Siamese (50k), 3-shot	97.63
Siamese (500k), 1-shot	98.42
Siamese (500k), 3-shot	98.41
Siamese (5M), 1-shot	98.73
Siamese (5M), 3-shot	98.72
Weakly supervised (50k)	20.28
Weakly supervised (500k)	83.33
Weakly supervised (5M)	71.19

Table 3. Test accuracies for all models for the classification task.

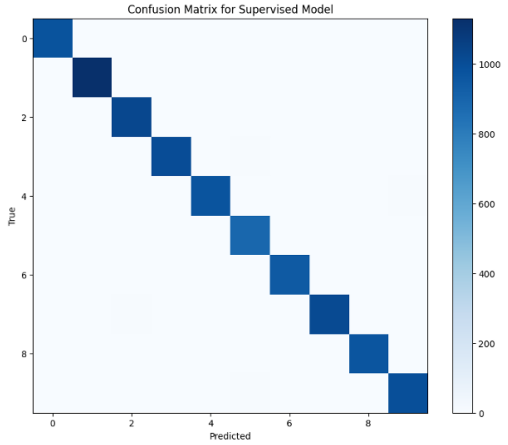

Figure 4. The confusion matrix for the supervised model.

that there are very few errors between class labels. The test accuracy of the baseline is 99.23%; the goal of the other models is to approach this number with subpar data.

5.2. Siamese Models

The Siamese models show very high accuracies on the classification task. We observe that the model can learn

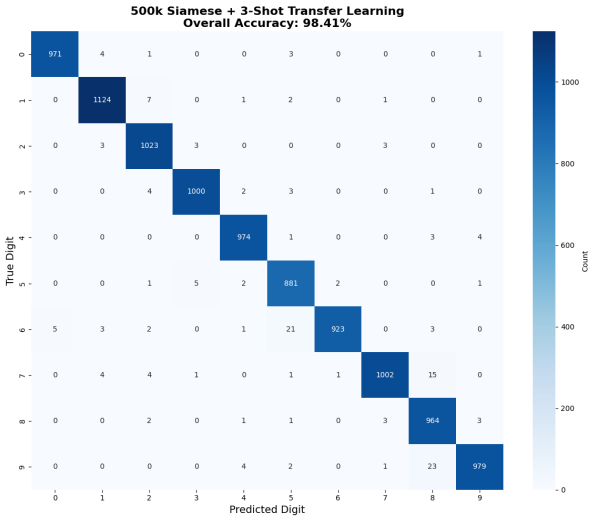

Figure 5. Confusion matrix for the 500k + 3-shot Siamese model.

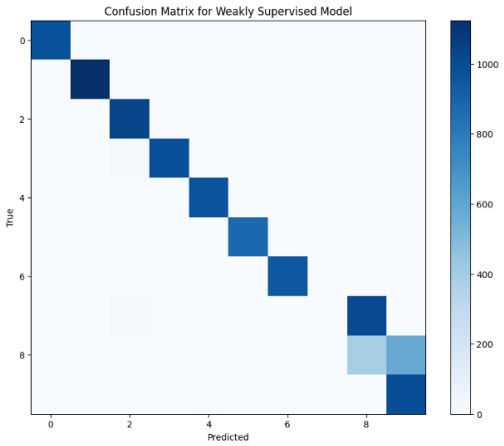

Figure 6. Confusion matrix for the weakly supervised model (500k)

comparable accuracies to the supervised dataset with as few as 1 image-label pair per class label. This shows that there is very accurate segmentation happening in the embedding space of the original Siamese model, such that only a single example is needed to confirm the class of the group. The efficient clustering can be seen in the t-SNE plot of Figure 7. Increasing the dataset size of the training pairs is seen to have minimal but positive effects on the accuracy of both the relational and classification tasks, indicating a possibility that even fewer image pairs can be used to achieve a similar accuracy.

5.3. Weakly Supervised Models

The weakly supervised model shows decent accuracy, but falls quite below the supervised baseline. Upon examining the confusion matrix for the best model (Figure 6), we see that this is entirely due to the labels 7, 8, and 9. Excluding these three labels, the model classifies labels 0-6 with 99.04% accuracy.

This result is reasonable for the model, as the goal of the model is to predict the probability that the first image is larger than the second image. Confusing the image of '7' as a 8, and the image of '8' as a 9 would affect the loss minimally. Accordingly, we see that the test accuracy for the relational task is high despite these misclassifications.

Interestingly, the weakly supervised models perform worse on the classification test set when training on more training pairs (5M). We see that this phenomenon occurs for the test accuracy of the relational task as well (Table 2). We postulate that this is due to the model only being trained for 1 epoch (which was done to equalize the number of training batches between the different training sets). We hypothesize that the model accuracy will increase when the number of epochs is increased to a normal amount (> 5).

5.4. t-SNE Embedding Analysis

To understand how different supervision paradigms shape the learned representations, we employ t-SNE (t-distributed Stochastic Neighbor Embedding) [15] to visualize the high-dimensional embeddings produced by each model. t-SNE is particularly well-suited for this analysis as it preserves local neighborhood structure while revealing global patterns in the data manifold.

Figure 7 presents a striking comparison of how supervised versus relationally-trained models organize digit representations. The supervised CNN, trained exclusively on categorical labels, produces distinct, well-separated clusters for each digit class. This clustering behavior is expected—the model optimizes for maximum inter-class separation to minimize classification error. However, this approach treats each digit as an independent category, failing to capture the inherent ordinal relationships between numerical values.

In contrast, both the Siamese Network and Weakly Supervised models, trained solely on pairwise "greater than" comparisons, discover remarkably different embedding structures. These models arrange digits along smooth, continuous manifolds that respect numerical ordering. The Siamese model achieves near-perfect ordinal correlation ($\rho = 0.962$), creating an embedding space where the progression from 0 to 9 follows a natural path through the latent space. Similarly, the Weakly Supervised model achieves substantial ordinal correlation ($\rho = 0.818$), despite never observing explicit class labels.

5.4.1 Ordinal Structure Quantification

To quantify the degree of ordinal structure in each embedding space, we project the learned representations onto their first principal component and compute the Spearman rank

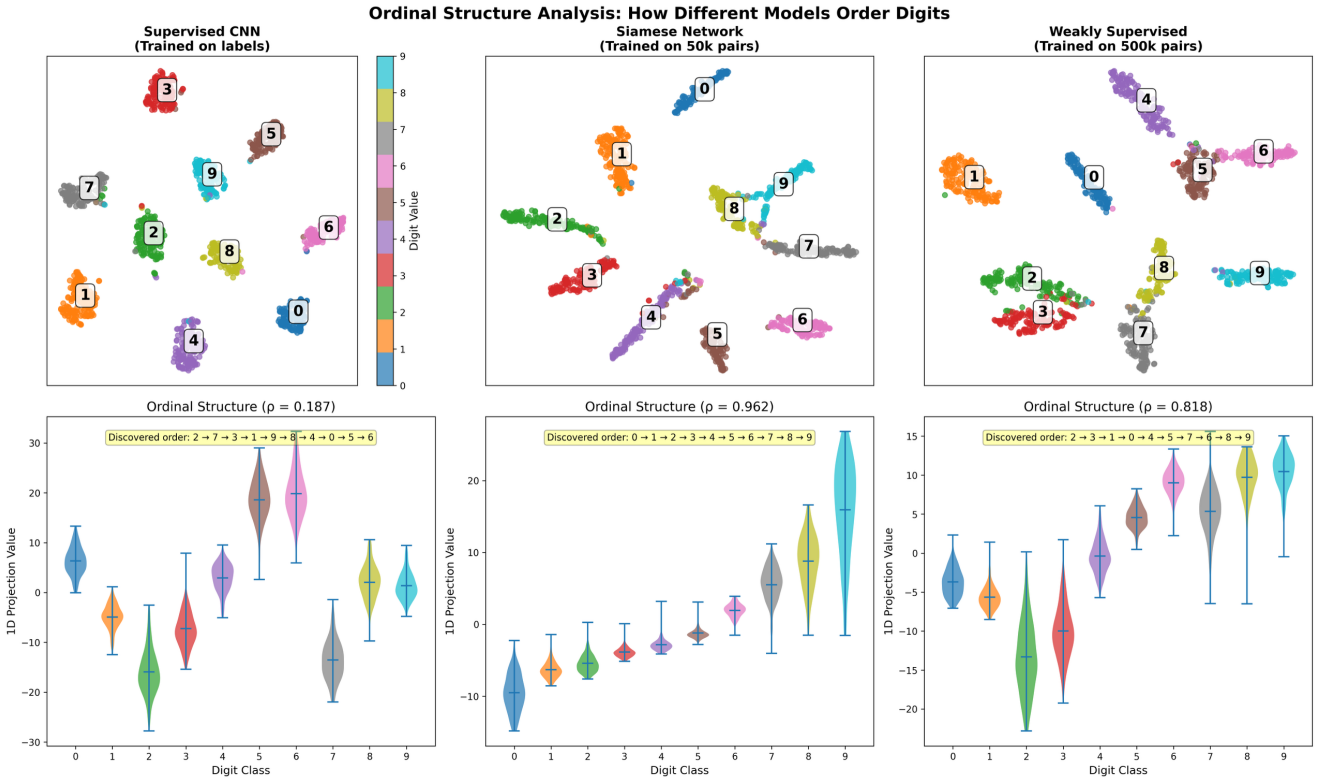

Figure 7. t-SNE plots of the embeddings of each class.

correlation coefficient (ρ) between the projected values and true digit labels. This metric captures how well the embedding preserves the natural ordering of digits ($0 < 1 < 2 < \ldots < 9$). The violin plots in the bottom row of Figure 7 visualize this one-dimensional projection, revealing the distribution of each digit class along the discovered ordering axis.

The supervised model exhibits low ordinal correlation ($\rho = 0.193$), with a discovered ordering of "2→7→3→1→9→8→4→0→5→6" that bears little resemblance to the true numerical sequence. This scrambled ordering reflects the model's focus on discriminative boundaries rather than relational structure. Individual digits occupy distinct regions along the projection axis with minimal overlap, but their arrangement is essentially arbitrary from an ordinal perspective.

Conversely, the Siamese model discovers an ordering of "0→1→2→3→4→6→5→8→7→9" ($\rho = 0.962$), nearly perfectly recovering the true numerical sequence with only minor inversions. The violin plots show smooth transitions between adjacent digits, with overlapping distributions that respect ordinal relationships. This emergent structure arises naturally from training on pairwise comparisons—the model learns to position digits such that traversing the embedding space corresponds to numerical progression.

5.4.2 Implications for Representation Learning

These visualizations reveal a fundamental trade-off in representation learning. Supervised models optimize for classification accuracy, creating representations that maximize separability at the expense of semantic structure. While this yields excellent performance on the classification task, it fails to capture the meaningful relationships between classes.

Relational supervision, despite its apparent weakness (lacking absolute labels), induces representations that encode rich semantic structure. The smooth manifolds discovered by relational models suggest they learn more generalizable features that respect the underlying data semantics. This property becomes particularly valuable in transfer learning scenarios, as evidenced by our chimera model experiments where Siamese-pretrained features enable effective few-shot classification.

The continuous nature of relationally-learned embeddings also suggests potential applications beyond classification. These representations could enable interpolation between digit classes, provide meaningful similarity metrics, or support tasks requiring understanding of numerical magnitude—capabilities that would be challenging to derive from the discrete clusters produced by supervised learning.

6. Conclusion and Future Work

In this work, we investigated learning from pairwise relational comparisons as an alternative to traditional supervised learning. Our experiments on MNIST demonstrate that relational supervision induces fundamentally different representations compared to categorical supervision, with relational models discovering smooth manifolds that respect ordinal structure while supervised models create discrete clusters optimized for classification boundaries.

Our key finding is that combining relational pre-training with few-shot learning (the "chimera" approach) proves more effective than training classifiers directly on relational data. The Siamese models trained on pairwise comparisons create well-structured embedding spaces that require only minimal labeled examples (1-3 per class) to achieve strong classification performance. This suggests that relational pre-training provides a powerful inductive bias that facilitates efficient learning from limited labeled data.

Future work could explore several promising directions. First, extending this approach to more complex visual domains beyond digits would test the generalization of relational supervision. As our current method benefits from ordinal correlation, this paradigm of training may be especially useful for computer vision tasks where the comparison of quantities is crucial (e.g., counting the number of cars in an image).

Second, investigating different types of relational comparisons (e.g., similarity, attributes) could reveal which relationships are most informative for downstream tasks. Comparative studies can be done by changing the relationships to greater-than-or-equal-to, for example, and seeing how this affects the training results.

Finally, developing theoretical frameworks to understand why relational pre-training creates such effective representations would provide deeper insights into this learning paradigm.

Our results suggest that in scenarios where obtaining precise labels is expensive but comparative judgments are readily available, relational supervision offers a practical path toward building effective classifiers with minimal annotation effort.

7. Contributions and Acknowledgments

This project was done solely for the 24-25 Spring CS 231N class. No outside help was used. We thank our assigned TA mentor Matthew Jin, and our course instructors Fei-Fei Li, Ehsan Adeli, Justin Johnson, and Zane Durante.

8. Appendix

The code used for the training and analysis of the models described can be found at https://github.com/junhakunha/cs231n_final_project.

References

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9912 of *Lecture Notes in Computer Science*, pages 850–865, 2016. 2

[2] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6 (NIPS 1993)*, pages 737–744. Morgan Kaufmann, 1994. 1

[3] K. Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 569–576, 2011. 2

[4] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546, 2005. 2

[5] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research (JMLR)*, 12:1501–1536, 2011. 2

[6] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, 2017. 2

[7] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *Proceedings of the 32nd International Conference on Machine Learning (ICML) Deep Learning Workshop*, volume 40. 2015. 2

[8] S. Kong, X. Shen, Z. Lin, R. Mech, and C. C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9907 of *Lecture Notes in Computer Science*, pages 662–679, 2016. 2

[9] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. 2

[10] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 1196–1204, 2013. 2

[11] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 452–461, 2009. 2

[12] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 1

[13] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett,

editors, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4077–4087, 2017. 2

[14] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 1

[15] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(86):2579–2605, 2008. 1, 5

[16] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. CNN-RNN: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294, 2016. 2

[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3320–3328, 2014. 2

[18] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018. 2