Conditional Contrastive Networks

Emily Mu
Massachusetts Institute of Technology
emilymu@mit.edu

John Guttag
Massachusetts Institute of Technology
guttag@csail.mit.edu

Abstract

A vast amount of structured information associated with unstructured data, such as images or text, is stored online. This structured information implies different similarity relationships among unstructured data. Recently, embeddings trained using contrastive learning on web-scraped unstructured data have been shown to have state-of-the-art performance across computer vision tasks. However, contrastive learning methods typically use only a single metric of similarity. In this paper, we propose conditional contrastive networks (CCNs) as a way of using multiple notions of similarity in structured data. Our novel conditional contrastive loss is able to learn multiple disjoint similarity notions by projecting each similarity notion into a different subspace. We show empirically that our CCNs perform better than single-label trained cross-entropy networks, single-label trained supervised-contrastive networks, multi-task trained cross-entropy networks, and previously proposed conditional similarity networks–on both the attributes on which it was trained and on unseen attributes.

1 Introduction

Structured tabular data associated with unstructured data, such as images and text, is a common method of data storage [Cafarella et al., 2008, Deng et al., 2022, Bordes et al., 2013]. Often, the structured tabular data captures important similarity relationships among the unstructured data entries. Some examples include patient data associated with chest radiographs [Irvin et al., 2019, Johnson et al., 2020] and relational tables associated with website text [Chen et al., 2000, Bhagavatula et al., 2015]. One example is given in Figure 1. Each of these images of shoes are associated with distinct category, closure, and gender attributes. Each of these attributes can define a metric of similarity between images.

ID	Category	Closure	Gender
1	Shoe	Lace up	Men
2	Shoe	Slip on	Women
3	Boot	Lace up	Women

Figure 1: **Shoe Example.** An example illustrating how structured tabular data can encode different similarity relationships between three images of shoes.

Contrastive learned embeddings trained on web-scraped unstructured data have been shown to have state-of-the-art performance on a variety of computer image tasks [Radford et al., 2021, Yuan et al., 2021, Khosla et al., 2020]. However, neither supervised nor unsupervised contrastive learning methods are able to leverage multiple metrics of similarity embedded in associated structured data. In unsupervised contrastive learning, representations are trained to discriminate pairs of similar

Table Representation Learning Workshop at NeurIPS 2022.

images (positive examples) from a set of dissimilar images (negative examples). Similar images are generated using label-preserving augmentations (e.g., rotations, changing brightness) [Chen et al., 2020]. Instead of relying on label-preserving augmentations, supervised contrastive learning approaches consider all instances with the same label to be positive examples [Khosla et al., 2020]. Neither unsupervised nor supervised contrastive learning are able to leverage different metrics of similarity where two examples may be similar under one metric and different under another. In this work, we propose conditional contrastive networks (CCNs). Our framework is shown in Figure 2. CCNs leverage multiple projection heads to learn embeddings from multiple metrics of similarity. In this way, we are able to represent examples that may be positive examples in one projected subspace and negative examples in a different projected subspace.

We compare our method to several existing methods that learn embeddings with different metrics of similarity. One method to learn many different metrics of similarity is to learn separate networks for each metric [Khosla et al., 2020]. However, this requires many parameters and multiple models, resulting in redundant embedding spaces that are not easily combined. A second method is to learn many different metrics of similarity is through multi-task learning [Ruder, 2017]. Multitask learning is typically used to leverage information of training signals of different tasks across the same domain. However, performing multitask learning on heterogeneous tasks can potentially hurt learning. The most similar method to ours is conditional similarity networks [Veit et al., 2017]. In conditional similarity networks, weights are learned or assigned to different embedding dimensions with respect to different metrics of similarity. These weights are learned jointly with the convolutional neural network parameters during training time. Conditional similarity networks are trained using triplets selected through similar and dissimilar categorical variables. However, this triplet selection process strictly reduces the amount of information available from the original full attribute data. This is because each triplet can only capture single attribute information between three data examples. Our conditional contrastive networks are able to leverage information that is not available using conditional similarity networks in order to improve embedding performance. In our experiments, we find that embeddings learned with CCNs outperform embeddings learned from both single label trained networks, multi-task trained networks, and conditional similarity networks on both in-domain and out-of-domain downstream tasks. Our main contributions are:

1. We propose conditional contrastive networks to utilize supervision from multiple different categorical notions of similarity.
2. We empirically demonstrate that a conditional contrastive network performs better than each of individually trained cross-entropy and supervised-contrastive networks supervised with a single notion of similarity, a multi-task trained cross-entropy networks, and a conditional similarity network.
3. We empirically demonstrate that the learned representation is more generalizable to out-of-distribution categories than any of the other learned representations.

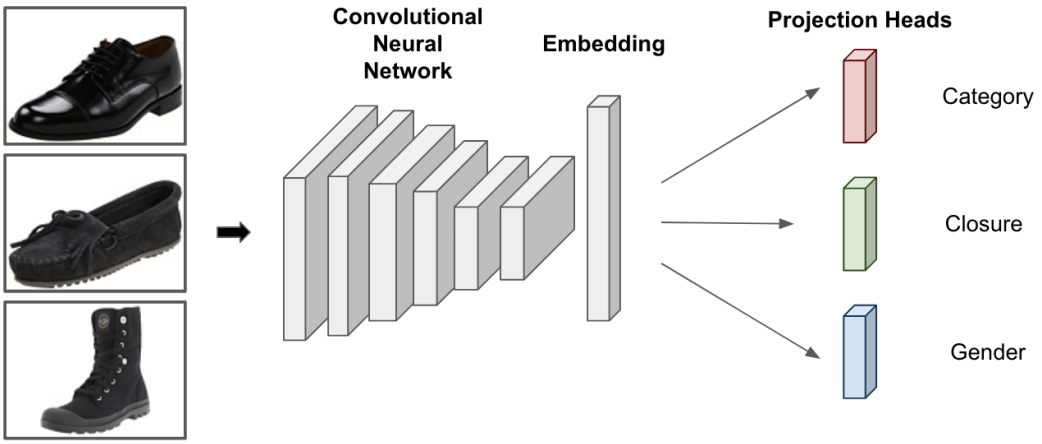

Figure 2: **Conditional Contrastive Network.** Multiple projection heads are trained to learn from multiple metrics of similarity. Both the base encoding network and the projection heads are trained together. The projection heads are discarded and only the encoding network is kept for downstream tasks.

2 Conditional Contrastive Network

Multi-task Setup We assume that during training time, we have access to dataset: $\mathcal{D}_1 = \{x_i, y_i^1, ...y_i^c\}_i^M$, where x is the unstructured data (e.g., image pixels) and the ys are the multiple categorical attributes available from the tabular data. We aim to learn an embedding function $f : X \rightarrow \mathbb{R}^d$ to map input data x to the representation space. We define $h_i = f(x_i)$ to be the embedding of x_i.

During contrastive training, we select a batch of N randomly sampled data/ $\{x_i\}_{i=1...N}$. We randomly sample 2 distinct augmentations for x_i, $\tilde{x}_{2i}$, and $\tilde{x}_{2i-1}$ to construct $2N$ samples, $\{\tilde{x}_j\}_{j=1...2N}$. We assume that all augmentations are label-preserving. We let $A(i) = \{1, ...2N\}\backslash i$ be the set of all samples not including i.

SimCLR A simple framework for contrastive learning representations (SimCLR) represents the unsupervised loss as follows [Chen et al., 2020]. g is defined as a projection head that maps the embedding to the surface of the unit sphere $\mathbb{S}^d = \{v \in \mathbb{R}^d : ||v||_2 = 1\}$. We define $v_i = g(h_i)$ to be the mapping of h_i on the unit sphere.

$$L^{simclr} = \sum_{i=I} -\log \frac{\exp(\frac{v_i^T v_p}{\tau})}{\sum_{a \in A(i)} \exp(\frac{v_i^T v_a}{\tau})}$$

$\tau \in \{0, \infty\}$ is the temperature hyperparameter. The positive example for each sample consists of transformed version of that sample. All other samples are considered to be negative samples. The SimCLR objective achieves state-of-the-art performance for unsupervised learning methods.

SupCon In contrast to SimCLR, supervised contrastive (SupCon) learning considers all samples with the same label to be positive examples for a given reference. SupCon loss is as follows [Khosla et al., 2020].

$$L^{supcon} = \sum_{i=I} L_i^{supcon} = \sum_{i=I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\frac{v_i^T v_p}{\tau})}{\sum_{a \in A(i)} \exp(\frac{v_i^T v_a}{\tau})}$$

where $|S|$ denotes the cardinality of the set S, $P(i)$ is defined as the positive set with all samples with the same label as x_i distinct from itself i.e., $P(i) = \{j \in A(i) : y_j = y_i\}$. Similarly, we define the negative set $N(i)$ to be all samples with a different label from x_i, $N(i) = \{j \in A(i) : y_j \neq y_i\}$. Although supervised contrastive learning is able to learn from labels, neither supervised nor unsupervised contrastive learning are able to consider multiple metrics of similarity.

CondCon We train the conditional contrastive network with our conditional contrastive loss (CondCon). We define separate projection heads g^c that maps the embedding to the surface of the unit sphere. We define $v_i^c = g^c(h_i)$ to be the mapping of h_i on the unit sphere by projection head g^c.

$$L^{condcon} = \sum_{c \in C} \sum_{i=I} L_{c,i}^{condcon} = \sum_{c \in C} \sum_{i=I} \frac{-1}{|P^c(i)|} \sum_{p \in P^c(i)} \log \frac{\exp(\frac{v_i^{cT} v_p^c}{\tau})}{\sum_{a \in A(i)} \exp(\frac{v_i^{cT} v_a^c}{\tau})}$$

$P^c(i)$ is defined as the positive set with the same label as x_i under attribute c.

3 Experiments

Dataset All of our experiments are performed on the Zappos50k dataset [Yu and Grauman, 2014, 2017]. This dataset contains 50,025 catalog images of single shoes collected from Zappos.com. This dataset also includes rich attribute information. We train our networks with shoe category (shoes, boots, sandals, slippers), closure mechanism (lace up, slip on, zipper, hook and loop, pull on), and gender (women, men, girls, boys). We also use the shoe brand information to evaluate model embeddings on out-of-domain classification tasks. We split the images into 3 parts: 70% for training, 10% for validation, and 20% for testing.

Models All models are trained with a ResNet18 backbone [He et al., 2016] pretrained on ImageNet [Deng et al., 2009] and an embedding space of 128. All images are resized to 112x112. We train using standard data augmentations, including random crops, flips, and color jitters. We train and evaluate the following networks

Table 1: **CCN trained embeddings have the highest performance across all evaluation settings.** We report the mean and the standard deviation of the test set classification accuracy.

	Evaluation Settings			
Model	Category	Closure	Gender	Brand (OOD)
XEnt Category	96.64 (0.61)	74.55 (1.77)	63.78 (2.11)	27.22 (1.18)
XEnt Closure	88.99 (1.47)	92.28 (1.16)	66.59 (1.55)	29.18 (0.58)
XEnt Gender	81.96 (1.58)	73.28 (1.62)	83.09 (2.31)	24.08 (0.38)
XEnt Multi	96.98 (0.92)	93.33 (1.37)	85.07 (1.15)	32.10 (0.74)
SupCon Category	96.95 (1.20)	73.02 (0.91)	61.24 (2.90)	27.87 (1.47)
SupCon Closure	83.62 (1.36)	91.75 (1.65)	65.90 (2.23)	26.78 (0.47)
SupCon Gender	76.40 (1.93)	69.52 (1.40)	85.11 (0.93)	24.30 (0.96)
CSN	83.33 (0.82)	72.12 (1.82)	69.21 (2.40)	16.27 (0.37)
CCN (Ours)	**97.30 (0.57)**	**94.26 (1.02)**	**86.38 (1.37)**	**43.49 (0.81)**

- **Cross-Entropy Networks (XEnt)** We train three cross-entropy networks with each of the three labels (category, closure, gender). We also train a multitask cross-entropy network trained with all three labels. We train each network for 200 epochs with a batch-size of 64 and a learning rate of 0.01.
- **Conditional Similarity Network (CSN)** We train a conditional similarity network that learns the convolutional filters, embedding, and mask parameters together. This network had the best performance out of all the conditional similarity variants. 10,000 triplets are constructed from the attributes for training (category, closure, gender) and the network is trained for 200 epochs. We follow the training procedure specified by Veit et al. [2017].
- **Supervised Contrastive Networks (SupCon)** We train three supervised contrastive networks with each of the three labels (category, closure, gender). All contrastive network projection heads have projection dimensions of 32. We train for 200 epochs with a batch-size of 64 and a learning rate of 0.05 with a stochastic gradient descent optimizer.
- **Conditional Contrastive Network (CCN)** We train a conditional contrastive network with three projection heads corresponding to the three training attributes. Hyperparameter settings are the same as for SupCon networks.

Results We evaluate all networks by training a softmax classifier layer over the frozen embedding. All networks are trained with a batch-size of 64 and a learning rate of 0.1 for 20 epochs. The checkpoint with the best validation accuracy is used to evaluate the test set.

For each model we evaluate on three in-domain classification tasks: category, closure, and gender. We also evaluate on one out-of-domain setting: classifying shoe brands. For the out-of-domain setting, we limit to only the top 20 brands. We report top-1 classification accuracy and standard deviation across 5 splits of the test set in Table 1.

The conditional contrastive network trained embeddings have the highest performance for all tasks. Furthermore, we find that CCN-trained embeddings outperform CSN-trained embeddings by a large margin. We hypothesize that this is because the conditional contrastive network is trained on strictly more information that the conditional similarity network. We also find that the greatest benefit of the conditional contrastive network in comparison to all other networks comes when generalizing to the out-of-domain classification task.

4 Conclusion

In this work, we propose conditional contrastive networks to leverage different metrics of similarity available in structured data. We demonstrate that our CCN-learned embeddings perform better than single-label trained cross-entropy embeddings, single-label trained supervised-contrastive embeddings, multi-task trained cross-entropy embeddings, and conditional similarity network embeddings on a suite of downstream tasks. The most striking improvement is the better generalization to to new tasks. One limitation of our work is that we only evaluate on a single dataset. We intend to validate our approach on a variety of domains.

References

Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Tabel: Entity linking in web tables. In *International Semantic Web Conference*, pages 425–441. Springer, 2015.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

Michael J Cafarella, Alon Y Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu. Uncovering the relational web. In *WebDB*, pages 1–6, 2008.

Hsin-Hsi Chen, Shih-Chung Tsai, and Jin-He Tsai. Mining tables from large scale html texts. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 830–838, 2017.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014.

Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579, 2017.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2 for contrastive framework set-up and assumptions.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will include this in the supplementary material and provide a GitHub link upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3 for some experimental details. Additional details will be available in the supplement.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report the mean as well as the standard deviation for all experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We will include this in the supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the original creators of all datasets in Section 3.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code will be available in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data we're using are open-source datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] .
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] .
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] .
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] .