Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

# Deep Metric Learning

---

20.07.2021

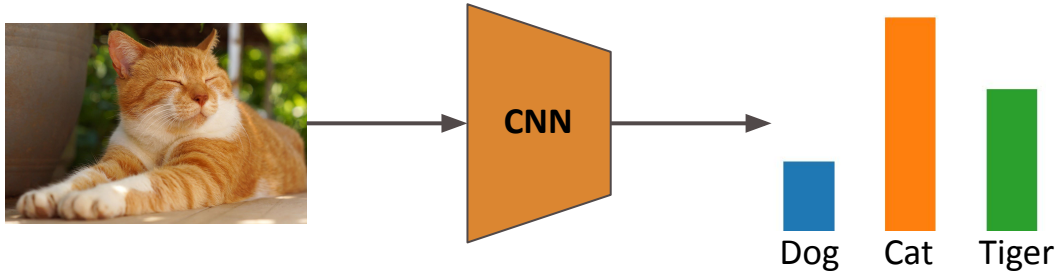
PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

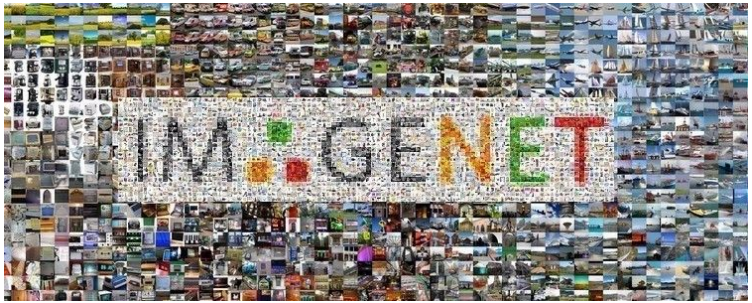
Contact: villar@ais.uni-bonn.de

# Motivation

# What can DL do for us?

## Classification

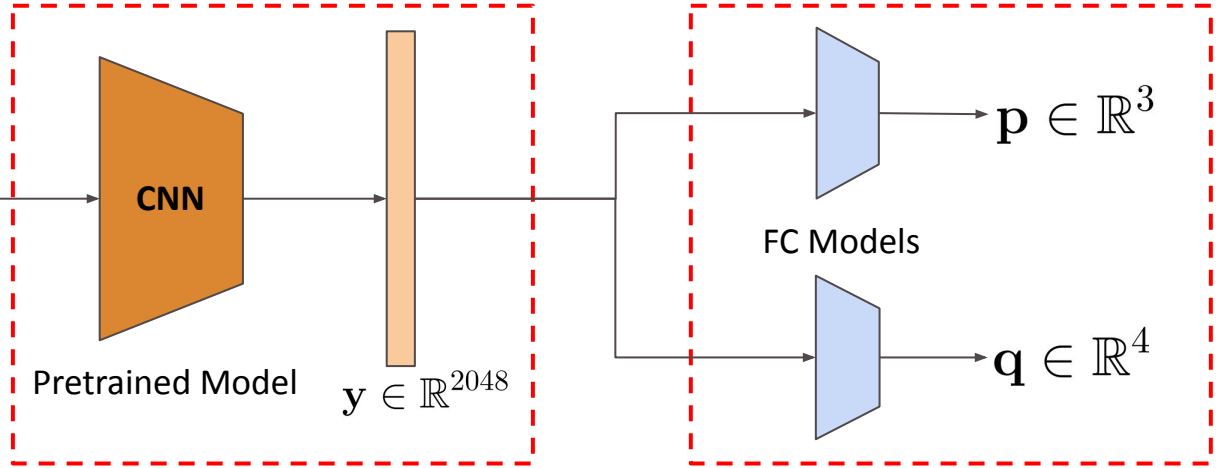



> 20,000 categories

# What can DL do for us?

**Regression**



# What can DL do for us?

**Third type of problems**

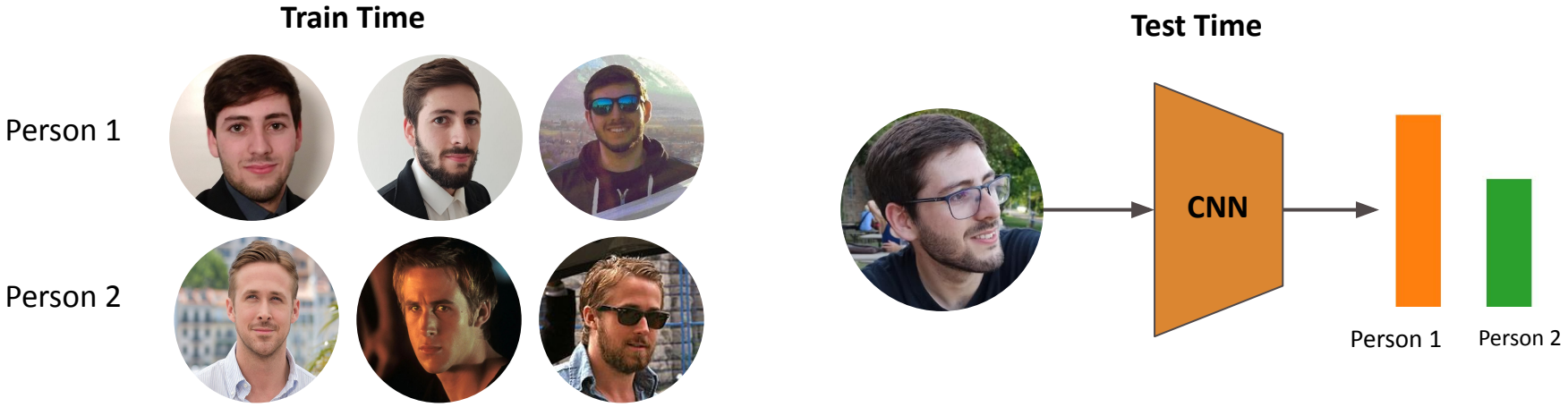
**Classification**: male, brown hair, suit

Is this the same person?

**Classification**: male, blond, shirt

# Why Similarity Learning?

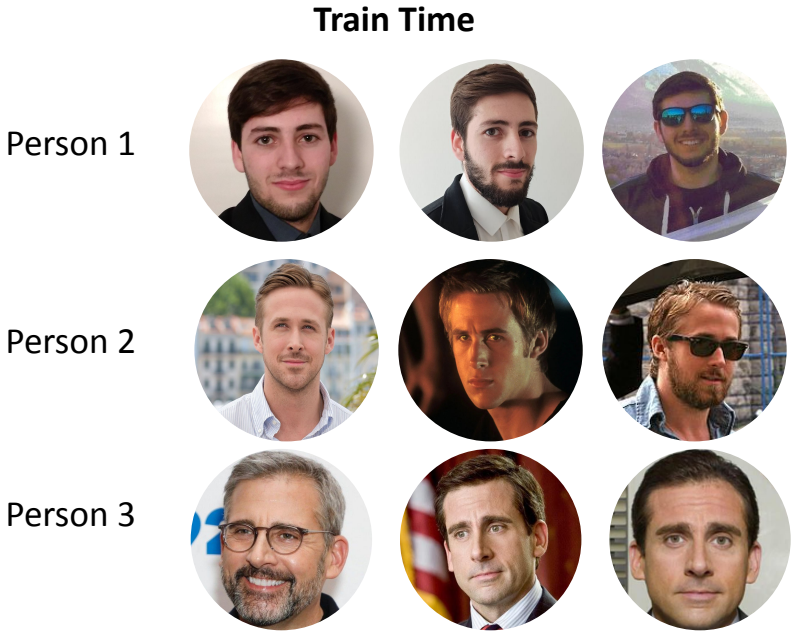
- Application: Face recognition to unlock a door

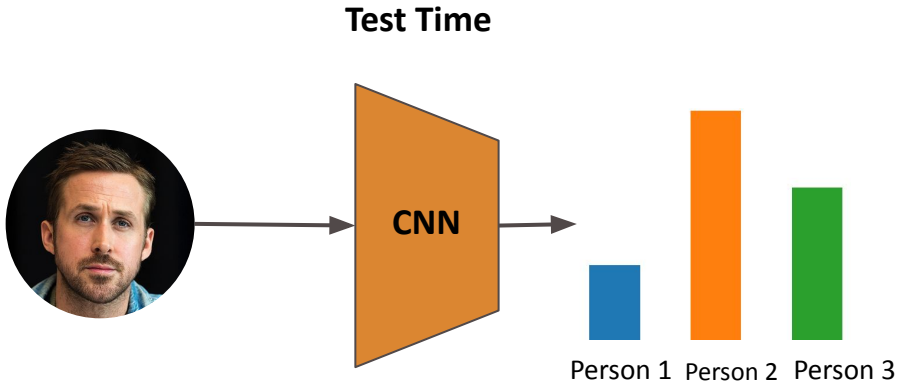



What is the problem with this approach?

# Why Similarity Learning?

- Application: Face recognition to unlock a door

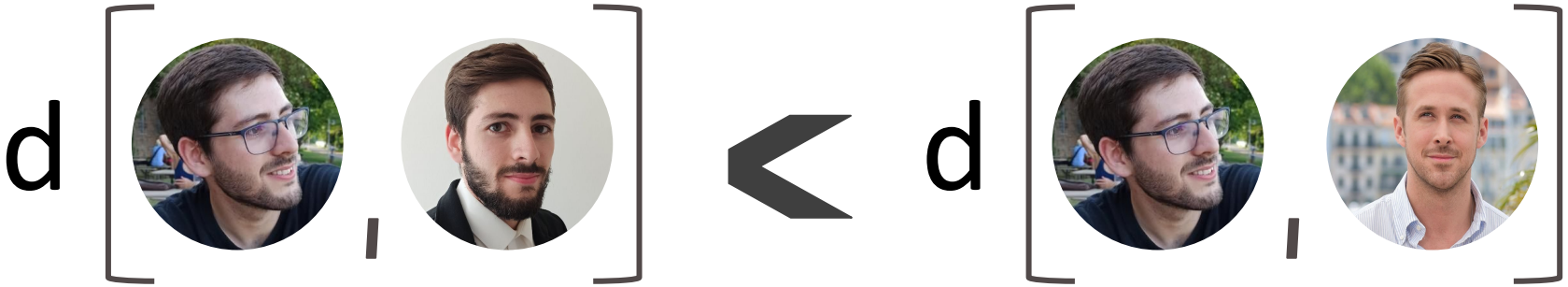



This is not scalable! We need to modify and retrain our model every time we add a new user

# Similarity Learning

- Learning a similarity/distance function




Small distance
High similarity

Large distance
Low similarity

# Similarity Learning

**Same Person:**

$$d(A, B) < \tau$$

$A$

$B$

**Different Person:**

$$d(A, C) > \tau$$

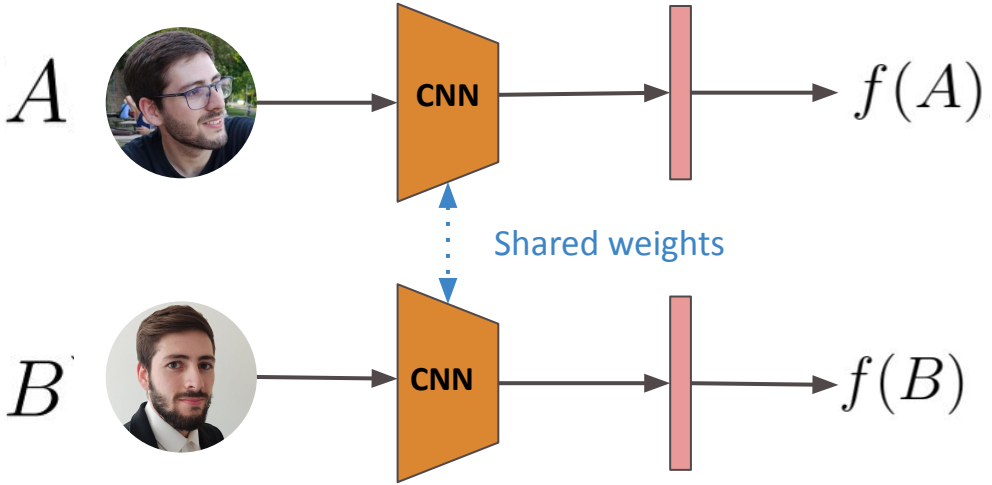
$A$

$C$

# Siamese Neural Networks

# Siamese Neural Networks

- *“Network that uses the same weights on two or more different inputs to compute comparable output vectors“*
- Output vectors of fixed dimensionality denoted as embeddings



# Similarity Learning

- Distance function:

$$d(A, B) = ||f(A) - f(B)||^2$$

- During training: learn model parameters such that:
  - $A$ and $B$ are the same person ⟹ $d(A, B)$ is small
  - $A$ and $B$ are a different person ⟹ $d(A, B)$ is large

# Contrastive Loss

- Loss function for a positive pair
  - $A$ and $B$ are the same person ⟹ $d(A, B)$ is small

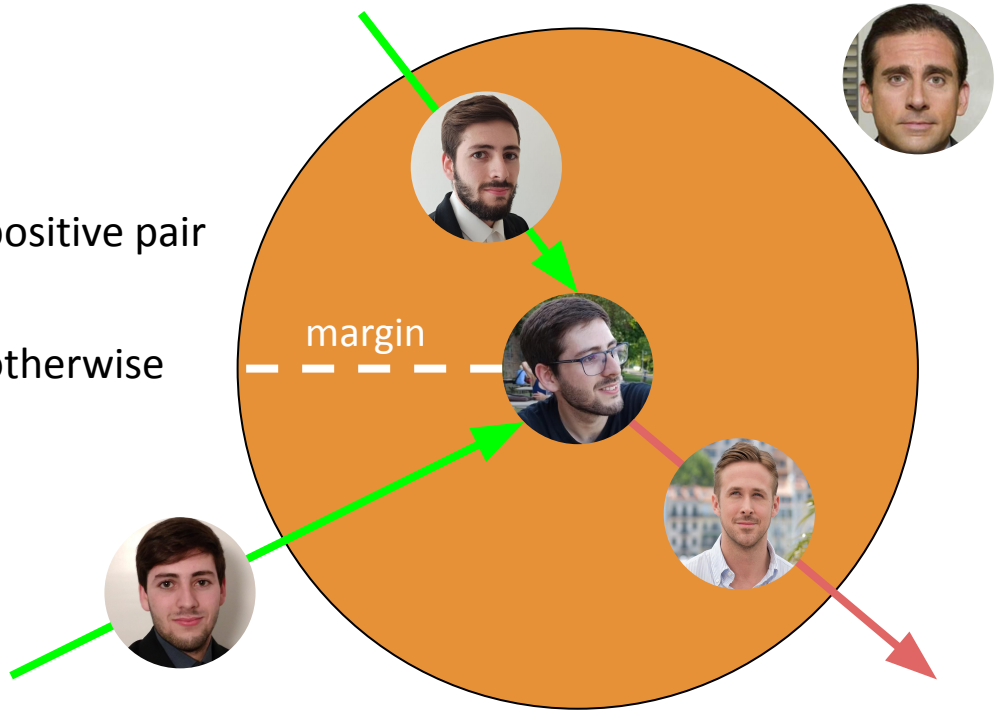
$$\mathcal{L}(A, B) = d(A, B)$$

- Loss function for a negative pair
  - $A$ and $B$ are a different person ⟹ $d(A, B)$ is large

$$\mathcal{L}(A, B) = max(0, \alpha - d(A, B))$$

# Contrastive Loss

$$\mathcal{L}(A,B) = \begin{cases} d(A,B) & \text{; positive pair} \\ max(0, \alpha - d(A,B)) & \text{; otherwise} \end{cases}$$



# Triplet Loss

- Works with triplets of images

Anchor (A)

Positive (P)

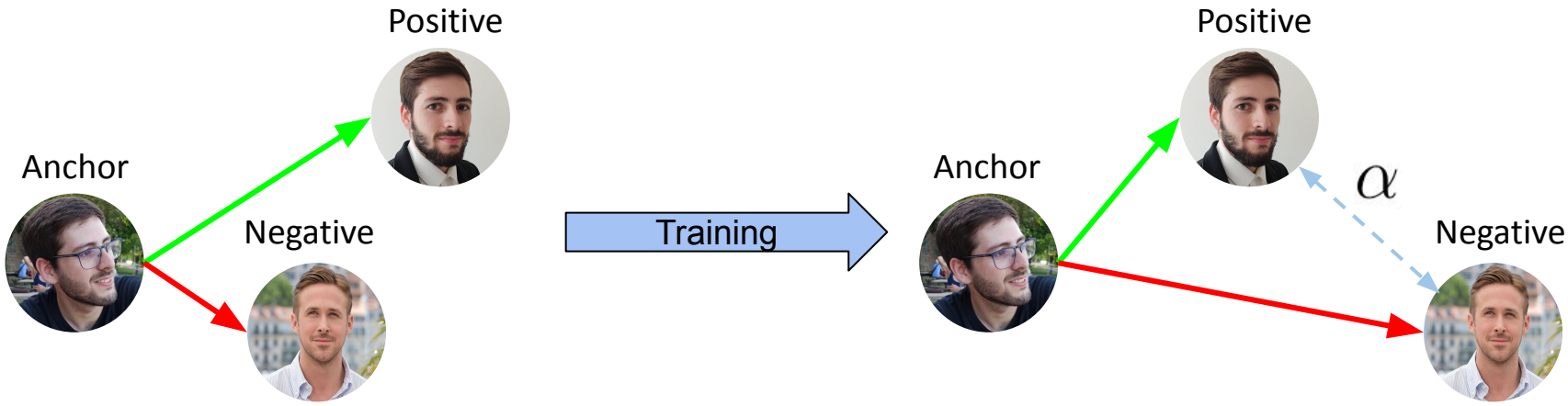
Negative (N)

- We want to obtain:

$$d(A, P) < d(A, N)$$

# Triplet Loss

$$\mathcal{L}(A, P, N) = \max\{0, \underbrace{d(A, P)}_{\text{minimize}} - \underbrace{d(A, N)}_{\text{maximize}} + \underbrace{\alpha}_{\text{Stop condition}}\}$$



# Applications and Results

# Person Recognition & Re-Identification

**FaceNet: A Unified Embedding for Face Recognition and Clustering**

Florian Schroff
fschroff@google.com
Google Inc.

Dmitry Kalenichenko
dkalenichenko@google.com
Google Inc.

James Philbin
jphilbin@google.com
Google Inc.

**Abstract**

*Despite significant recent advances in the field of face recognition [10, 14, 15, 17], implementing face verification and recognition efficiently at scale presents serious challenges to current approaches. In this paper we present a system, called FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Once this space has been produced, tasks such as face recognition, verification and clustering can be easily implemented using standard techniques with FaceNet embeddings as feature vectors.*

*Our method uses a deep convolutional network trained to directly optimize the embedding itself, rather than an intermediate bottleneck layer as in previous deep learning approaches. To train, we use triplets of roughly aligned matching / non-matching face patches generated using a novel online triplet mining method. The benefit of our approach is much greater representational efficiency: we achieve state-of-the-art face recognition performance using only 128-bytes per face.*

*On the widely used Labeled Faces in the Wild (LFW) dataset, our system achieves a new record accuracy of* ***99.63%****. On YouTube Faces DB it achieves* ***95.12%****. Our system cuts the error rate in comparison to the best published result [15] by 30% on both datasets.*

## 1. Introduction

In this paper we present a unified system for face verification (is this the same person), recognition (who is this person) and clustering (find common people among these faces). Our method is based on learning a Euclidean embedding per image using a deep convolutional network. The network is trained such that the squared L2 distances in the embedding space directly correspond to face similarity: faces of the same person have small distances and faces of

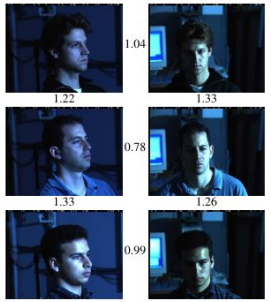



Figure 1. **Illumination and Pose invariance.** Pose and illumination have been a long standing problem in face recognition. This figure shows the output distances of FaceNet between pairs of faces of the same and a different person in different pose and illumination combinations. A distance of 0.0 means the faces are identical, 4.0 corresponds to the opposite spectrum, two different identities. You can see that a threshold of 1.1 would classify every pair correctly.

tion problem; and clustering can be achieved using off-the-shelf techniques such as k-means or agglomerative clustering.

Previous face recognition approaches based on deep networks use a classification layer [15, 17] trained over a set of known face identities and then take an intermediate bottleneck layer as a representation used to generalize recognition

**In Defense of the Triplet Loss for Person Re-Identification**

Alexander Hermans*, Lucas Beyer* and Bastian Leibe
Visual Computing Institute
RWTH Aachen University
last@vision.rwth-aachen.org



**Abstract**

*In the past few years, the field of computer vision has gone through a revolution fueled mainly by the advent of large datasets and the adoption of deep convolutional neural networks for end-to-end learning. The person re-identification subfield is no exception to this. Unfortunately, a prevailing belief in the community seems to be that the triplet loss is inferior to using surrogate losses (classification, verification) followed by a separate metric learning step. We show that, for models trained from scratch as well as pretrained ones, using a variant of the triplet loss to perform end-to-end deep metric learning outperforms most other published methods by a large margin.*

## 1. Introduction

In recent years, person re-identification (ReID) has attracted significant attention in the computer vision community. Especially with the rise of deep learning, many new approaches have been proposed to achieve this task [40, 8, 42, 31, 39, 10, 52, 4, 46, 20, 54, 35][1]. In many aspects person ReID is similar to image retrieval, where significant progress has been made and where deep learning has recently introduced a lot of changes. One prominent example in the recent literature is FaceNet [29], a convolutional neural network (CNN) used to learn an embedding for faces. The key component of FaceNet is to use the triplet loss, as introduced by Weinberger and Saul [41], for training the CNN as an embedding function. The triplet loss optimizes the embedding space such that data points with the same identity are closer to each other than those with different identities. A visualization of such an embedding is shown in Figure 1.

Several approaches for person ReID have already used some variant of the triplet loss to train their models [17, 9, 28, 8, 40, 31, 33, 26, 6, 25], with moderate success. The

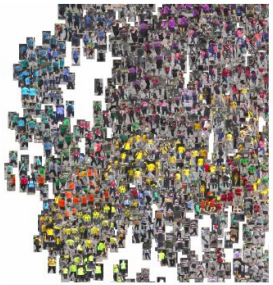


Figure 1: A small crop of the Barnes-Hut t-SNE [38] of our learned embeddings for the Market-1501 test-set. The triplet loss learns semantically meaningful features.

recently most successful person ReID approaches argue that a classification loss, possibly combined with a verification loss, is superior for the task [6, 51, 10, 52, 22]. Typically, these approaches train a deep CNN using one or multiple of these surrogate losses and subsequently use a part of the network as a feature extractor, combining it with a metric learning approach to generate final embeddings. Both of these losses have their problems, though. The classification loss necessitates a growing number of learnable parameters as the number of identities increases, most of which will be discarded after training. On the other hand, many of the networks trained with a verification loss have to be used

# Person Recognition & Re-Identification



# Content-Based Retrieval

**Learning Deep Representations of Medical Images using Siamese CNNs with Application to Content-Based Image Retrieval**

**Yu-An Chung  Wei-Hung Weng**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, Cambridge, MA 02139
{andyyuan,ckbjimmy}@mit.edu

**Abstract**

Deep neural networks have been investigated in learning latent representations of medical images, yet most of the studies limit their approach in a single supervised convolutional neural network (CNN), which usually rely heavily on a large scale annotated dataset for training. To learn image representations with less supervision involved, we propose a deep Siamese CNN (SCNN) architecture that can be trained with only binary image pair information. We evaluated the learned image representations on a task of content-based medical image retrieval using a publicly available multiclass diabetic retinopathy fundus image dataset. The experimental results show that our proposed deep SCNN is comparable to the state-of-the-art single supervised CNN, and requires much less supervision for training.

**1 Introduction**

Effective feature extraction and data representation are key factors to successful medical imaging tasks. Researchers usually adopt medical domain knowledge and ask for annotations from clinical experts. For example, using traditional image processing techniques such as filters or edge detection techniques to extract clinically relevant spatial features from images obtained by different image modalities, such as mammography [Tsochatzidis et al., 2017], lung computed tomography (CT) [Dhara et al., 2017], and brain magnetic resonance imaging (MRI) [Jenitta and Ravindran, 2017]. The handcrafted features with supervised learning using expert-annotated labels work appropriately for specific scenarios. However, using predefined expert-derived features for data representation limits the chance to discover novel features. It is also very expensive to have clinicians and experts to label the data manually, and such labor-intensive annotation task limits the scalability of learning generalizable medical imaging representations.

**End-to-end Learning of Deep Visual Representations for Image Retrieval**

**Albert Gordo · Jon Almazán · Jerome Revaud · Diane Larlus**



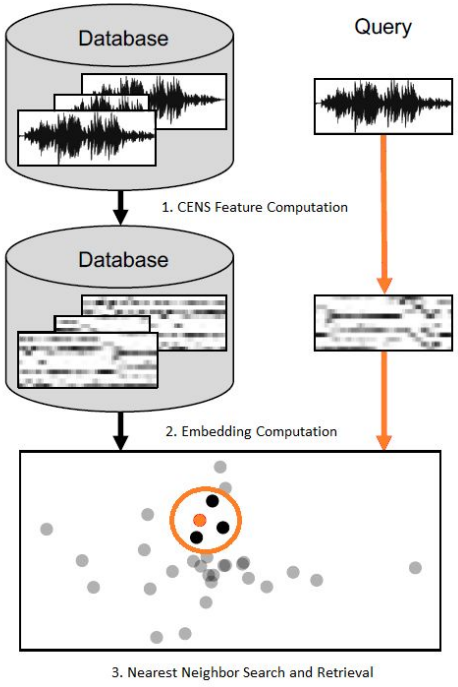
Received: date / Accepted: date

**Abstract** While deep learning has become a key ingredient in the top performing methods for many computer vision tasks, it has failed so far to bring similar improvements to instance-level image retrieval. In this article, we argue that reasons for the underwhelming results of deep methods on image retrieval are threefold: i) noisy training data, ii) inappropriate deep architecture, and iii) suboptimal training procedure. We address all three issues. First, we leverage a large-scale but noisy landmark dataset and develop an automatic cleaning method that produces a suitable training set for deep retrieval. Second, we build on the recent R-MAC descriptor, show that it can be interpreted as a deep and differentiable architecture, and present improvements to enhance it. Last, we train this network with a siamese architecture that combines three streams with a triplet loss. At the end of the training process, the proposed architecture produces a global image representation in a single forward pass that is well suited for image retrieval. Extensive experiments show that our approach significantly outperforms previous retrieval approaches, including state-of-the-art methods based on costly local descriptor indexing and spatial verification. On Oxford 5k, Paris 6k and Holidays, we respectively report 94.7, 96.6, and 94.8 mean average precision. Our representations can also be heavily compressed using product quantization with little loss in accuracy. To ensure the reproducibility of our research we have also released the clean annotations of the dataset and our pretrained models: http://www.xrce.xerox.com/Deep-Image-Retrieval.

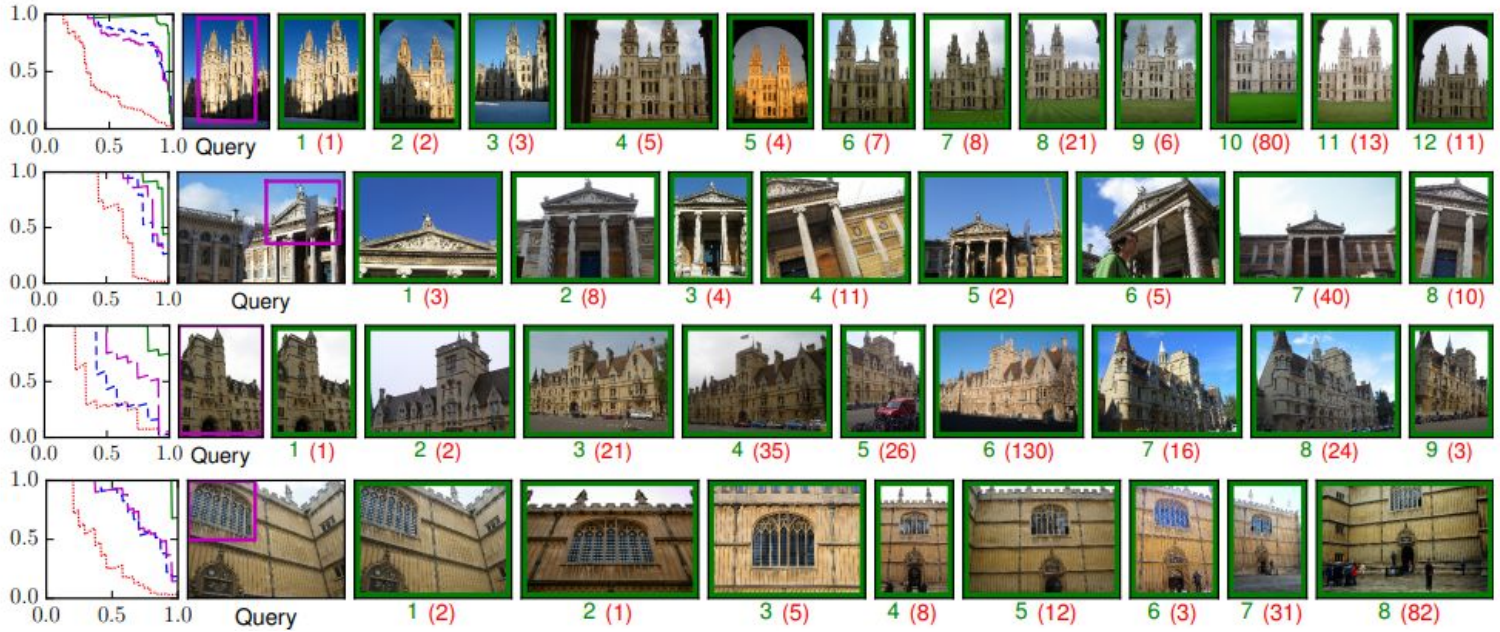
**1 Introduction**

*Instance-level image retrieval* is a visual search task that aims at, given a query image, retrieving all images that contain the same object instance as the query within a potentially very large database of images. Image retrieval and other related visual search tasks have a wide range of applications, *e.g.*, reverse image search on the web or organization of personal photo collections. Image retrieval has also been seen as a crucial component for data-driven methods that use visual search to transfer annotations associated with the retrieved images to the query image (Torralba et al, 2008). This has proved useful for annotations as diverse as image-level tags (Makadia et al, 2008), GPS coordinates (Hays and Efros, 2008), or prominent object location (Rodriguez-Serrano et al, 2015).

Deep learning, and particularly deep convolutional neural networks (CNN), have become an extremely powerful tool in computer vision. After Krizhevsky et al (2012) achieved the first place on the ImageNet classification and localization challenges in 2012 (Russakovsky et al, 2015) using a convolutional neural network, deep learning-based methods have significantly improved the state of the art in other tasks such as object detection (Girshick et al, 2014) and semantic segmentation (Long et al, 2015). Recently, they have also shined in other semantic tasks such as image captioning (Frome et al, 2013; Karpathy et al, 2014) and visual question answering (Antol et al, 2015). However, deep learning has been less successful so far in instance-level image



# Content-Based Retrieval



# Self-Supervised Pretraining

**Momentum Contrast for Unsupervised Visual Representation Learning**

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

Code: https://github.com/facebookresearch/moco

**Abstract**

*We present Momentum Contrast (MoCo) for unsupervised visual representation learning. From a perspective on contrastive learning [29] as dictionary look-up, we build a dynamic dictionary with a queue and a moving-averaged encoder. This enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. MoCo provides competitive results under the common linear protocol on ImageNet classification. More importantly, the representations learned by MoCo transfer well to downstream tasks. MoCo can* ***outperform*** *its supervised pre-training counterpart in* **7** *detection/segmentation tasks on PASCAL VOC, COCO, and other datasets, sometimes surpassing it by large margins. This suggests that the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks.*

## 1. Introduction

Unsupervised representation learning is highly successful in natural language processing, *e.g.*, as shown by GPT [50, 51] and BERT [12]. But supervised pre-training is still dominant in computer vision, where unsupervised methods generally lag behind. The reason may stem from differences in their respective signal spaces. Language tasks have discrete signal spaces (words, sub-word units, *etc.*) for building tokenized *dictionaries*, on which unsupervised learning can be based. Computer vision, in contrast, further concerns dictionary building [54, 9, 5], as the raw signal is in a continuous, high-dimensional space and is not structured for human communication (*e.g.*, unlike words).

Several recent studies [61, 46, 36, 66, 35, 56, 2] present promising results on unsupervised visual representation learning using approaches related to the *contrastive loss* [29]. Though driven by various motivations, these methods can be thought of as building *dynamic dictionaries*. The

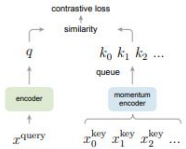



Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query $q$ to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, \ldots\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the current mini-batch enqueued and the oldest mini-batch dequeued, decoupling it from the mini-batch size. The keys are encoded by a slowly progressing encoder, driven by a momentum update with the query encoder. This method enables a large and consistent dictionary for learning visual representations.

From this perspective, we hypothesize that it is desirable to build dictionaries that are: (i) large *and* (ii) consistent as they evolve during training. Intuitively, a larger dictionary may better sample the underlying continuous, high-dimensional visual space, while the keys in the dictionary should be represented by the same or similar encoder so that their comparisons to the query are consistent. However, existing methods that use contrastive losses can be limited in one of these two aspects (discussed later in context).

We present *Momentum Contrast* (MoCo) as a way of building large and consistent dictionaries for unsupervised learning with a contrastive loss (Figure 1). We maintain the dictionary as a *queue* of data samples: the encoded representations of the current mini-batch are enqueued, and the

**A Simple Framework for Contrastive Learning of Visual Representations**

Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

**Abstract**

This paper presents *SimCLR*: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 85.8% top-5 accuracy, outperforming AlexNet with 100× fewer labels. [1]

## 1. Introduction

Learning effective visual representations without human

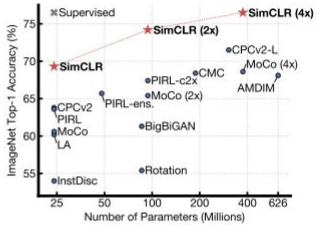



*Figure 1.* ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.

However, pixel-level generation is computationally expensive and may not be necessary for representation learning. Discriminative approaches learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset. Many such approaches have relied on heuristics to design pretext tasks (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018), which could limit the generality of the learned representations. Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-art results (Hadsell et al., 2006; Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019).

# References


1. https://towardsdatascience.com/all-you-want-to-know-about-deep-learning-8d68dcffc258
2. Goodfellow, Ian, et al. Deep learning. Vol. 1. No. 2. Cambridge: MIT press, 2016.
3. Advanced Deep Learning for Computer Vision, Lecture Notes. https://www.youtube.com/watch?v=6e65XfwmIWE
4. Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 2. IEEE, 2006.
5. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
6. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.
7. He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

# References

8. Hermans, Alexander, Lucas Beyer, and Bastian Leibe. "In defense of the triplet loss for person re-identification." arXiv preprint arXiv:1703.07737 (2017).

9. Gordo, Albert, et al. "End-to-end learning of deep visual representations for image retrieval." International Journal of Computer Vision 124.2 (2017): 237-254.