Sumer School XR&AI Matera, Italy

ellis

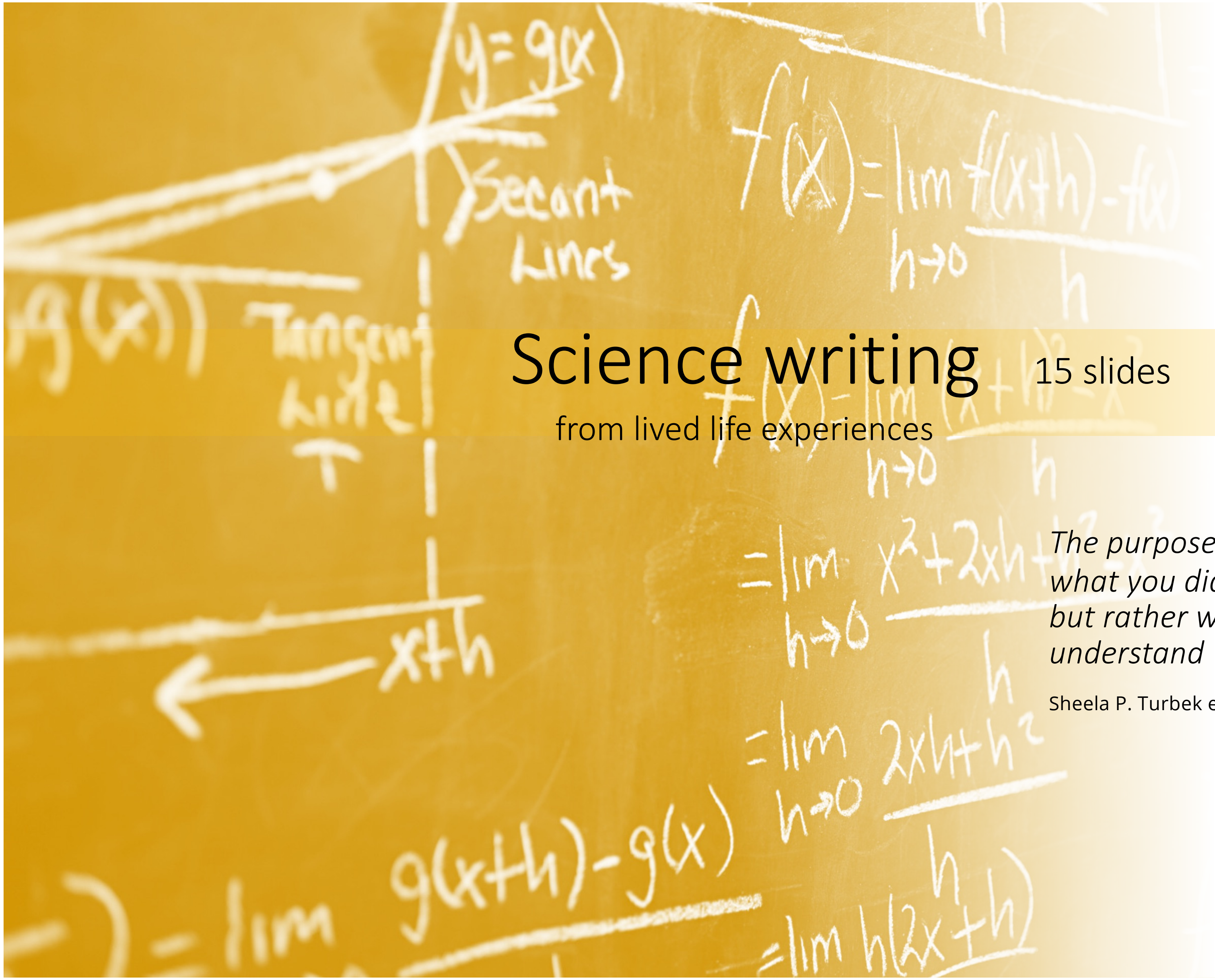
European Laboratory for Learning and Intelligent Systems

ELLIS Summer school

Large-scale Artificial Intelligence

Modena, 18-22 September 2023

UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS 1175
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA



Science writing

15 slides

from lived life experiences

Alberto Del Bimbo
Università di Firenze

The purpose of science writing is not explaining what you did or what you learned, but rather what you want your audience to understand

Sheela P. Turbek et al. , Ecology 101, 2016

IEEE Transactions on Pattern Analysis and Machine Intelligence
2023/3/20

1

CoReS: Compatible Representations via Stationarity

Niccolò Biondi, Federico Pernici, Matteo Bruni, and Alberto Del Bimbo, *Senior Member, IEEE*

Abstract—Compatible features enable the direct comparison of old and new learned features allowing to use them interchangeably over time. In visual search systems, this eliminates the need to extract new features from the gallery-set when the representation model is upgraded with novel data. This has a big value in real applications as re-indexing the gallery-set can be computationally expensive when the gallery-set is large, or even infeasible due to privacy or other concerns of the application. In this paper, we propose CoReS, a new training procedure to learn representations that are *compatible* with those previously learned, grounding on the stationarity of the features as provided by fixed classifiers based on polytopes. With this solution, classes are maximally separated in the representation space and maintain their spatial configuration stationary as new classes are added, so that there is no need to learn any mappings between representations nor to impose pairwise training with the previously learned model. We demonstrate that our training procedure largely outperforms the current state of the art and is particularly effective in the case of multiple upgrades of the training-set, which is the typical case in real applications.

Index Terms—Deep Convolutional Neural Network, Representation Learning, Compatible Learning, Fixed Classifiers.

ABSTRACT 191
1 INTRODUCTION 668
2 RELATED WORKS
- **Compatible Representation Learning.** 946
- **Neural Collapse.**
- **Class-incremental Learning.**

3 COMPATIBILITY EVALUATION 844
- **3.1 Compatibility Criterion**
- **3.2 Multi-step Compatibility Criterion**

4 LEARNING COMPATIBLE REPRESENTATIONS
- **4.1 Learning Stationary Features with Fixed Classifiers** 1710
- **4.2 Learning Compatibility via Stationarity**
- **4.3 CoReS Extension by Model Selection**

5 EXPERIMENTAL RESULTS
- **5.1 Compatibility Evaluation on CIFAR-100/10**
- **5.2 Compatibility Evaluation on CASIA-WebFace/LFW** 2403
- **5.3 Compatibility Evaluation on Market1501**
- **5.4 Compatibility Evaluation on GLDv2**
- **5.5 Compatibility Evaluation on MET**
- **5.6 Qualitative Results**

6 ABLATION STUDIES
- **6.1 Number of Epochs**
- **6.2 Model Selection** 1710
- **6.3 Pre-training**
- **6.4 Model Initialization**
- **6.5 Different class order**
- **6.6 Different Model Architectures**
- **6.7 Different Number of Future Classes**
- **6.8 Different Source of Upgrade Data**

7 CONCLUSIONS 193
ACKNOWLEDGMENTS
REFERENCES #72

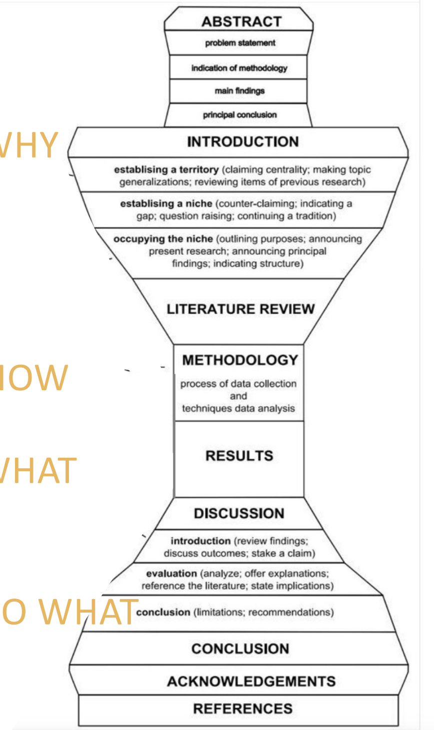
WHY

HOW

WHAT

SO WHAT

THE ANATOMY OF
A RESEARCH PAPER



Asad Naveed
Published in Linkedin
2023

Title

The title is the calling card of the work

Avoid:

- Non-significant words and phrases (e.g. *basically, kind of, actually, furthermore…*
- Unnecessary prepositions or articles (*e.g. due to the fact that, the field of, for the purpose of, that, the…)*
- Introductory phrases (e.g., *a study of*…)

The title should:

- Include keywords that summarize the content of the work
- Not be too generic, but informative, and precise. Hinting at the conclusions of the work allows for an informative title
- Be concise
- Attract the reader's attention….Personally, I don't like titles that aim to surprise

Titles from CVPR2023

- *Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert*
- *VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation*
- *SuperDisco: Super-Class Discovery Improves Visual Recognition for the Long-Tail*
- *REVEAL: Retrieval-Augmented Visual-Language Pre-Training With Multi-Source Multimodal Knowledge Memory*

- *Seeing Beyond the Brain: Conditional Diffusion Model With Sparse Masked Modeling for Vision Decoding*
- *Seeing a Rose in Five Thousand Ways*

- *Deep Graph-Based Spatial Consistency for Robust Non-Rigid Point Cloud Registration*
- *Cascaded Local Implicit Transformer for Arbitrary-Scale Super-Resolution*
- *Data-driven Feature Tracking for Event Cameras*

Abstract and Conclusions

Highly important sections

- Abstract: write it last. Do an exercise of extreme synthesis. Write motivations and contributions of your research. Only most important elements of how you obtain your results

- Conclusions: write it just before the abstract. Do not make a report of the sections. Just summarize your achievements and add criticisms. Perspectives are important if any…Don’t be concise

Introduction

- Define the significance of the problem you intend to address
- Describe what are the deficiencies or controversies in the current literature
- Put a few references, just make an overview of SoA. Reference just the most recent literature, with the exception of those articles that represent milestones in the field. The reviewer doesn’t care if you read a lot of papers…
- Provide a clear and precise explanation of the aim of the study. Put it in the last paragraph…
- Don’t present, analyze, or discuss the results. You would reduce the reader's curiosity. Just highlight your contributions.

In the final version

1 Introduction

Natural intelligent systems learn from visual experience and seamlessly exploit such learned knowledge to identify similar entities. Modern artificial intelligence systems, on their turn, typically require distinct phases to perform such visual search. An internal representation is first learned from a set of images (the *training-set*) using Deep Convolutional Neural Network models (DCNNs) [1], [2], [3], [4] and then used to index a large corpus of images (the *gallery-set*). Finally, visual search is obtained by identifying the closest images in the gallery-set to an input *query-set* by comparing their representations. Successful applications of learning feature representations are: face-recognition [5], [6], [7], [8], [9], person re-identification [10], [11], [12], [13], image retrieval [14], [15], [16], and car re-identification [17] among the others.

In the case in which novel data for the training-set and/or more recent or powerful network architectures become available, the representation model may require to be *upgraded* to improve its search capabilities. In this case, not only the query-set but also all the images in the gallery-set should be re-processed by the upgraded model to generate new features and replace the old ones to benefit from such upgrading. The re-processing of the gallery-set is referred to as *re-indexing* (Fig. 1).

For visual search systems with a large gallery-set, such as in surveillance systems, social networks or in autonomous robotics, re-indexing is clearly computationally expensive [18] or has critical deployment, especially when the working system requires multiple upgrades or there are real-time constraints. Re-indexing all the images in the gallery-set can be also infeasible when, due to privacy or ethical concerns,

The authors are with Media Integration and Communication Center (MICC), Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Firenze, 50139, Firenze, Italy. E-mail: name.surname@unifi.it

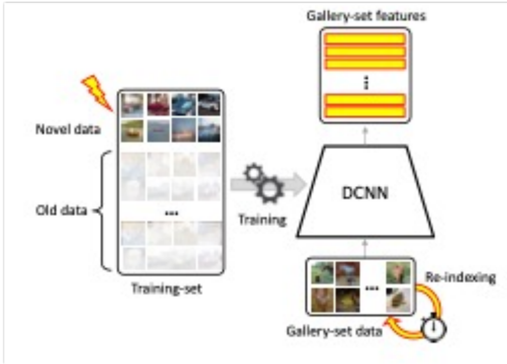


Fig. 1. Upgrading the DCNN representation model with novel data, typically requires the gallery-set to be re-indexed. Learning compatible representations allows to compare the newly learned representation of an input query-set with the old representation of the gallery-set, thus eliminating its computationally intensive re-indexing.

the original gallery images cannot be permanently stored [19] and the only viable solution is to continue using the feature vectors previously computed. In all these cases, it should be possible to directly compare the upgraded features of the query with the previously learned features of the gallery, i.e., the new representation should be *compatible* with the previously learned representation.

Learning compatible representation has recently received increasing attention and novel methods have been proposed in [18], [20], [21], [22], [23], [24]. Differently from these works, in this paper we address *compatibility* leveraging the *stationarity* of the learned internal representation. Stationarity allows to maintain the same distribution of the features over time so that it is possible to compare the features of the upgraded representation with those previously learned. In particular, we enforce stationarity by leveraging the properties of a family of classifiers whose parameters are not subject to learning, namely *fixed classifiers* based on regular polytopes [25] [26] [27], that allow to reserve regions of the representation space to future classes while classes already learned remain in the same spatial configuration.

The main contributions of our research are the following:

1) We identify stationarity as a key property for compatibility and propose a novel training procedure for learning compatible feature representations via stationarity, without the need of learning any mappings between representations nor to impose pairwise training with the previously learned model. We called our method: Compatible Representations via Stationarity (CoReS).
2) We introduce new criteria for comparing and evaluating compatible representations in the case of sequential multi-model upgrading.
3) We demonstrate through extensive evaluation on large scale verification, re-identification and retrieval benchmarks that CoReS improves the current state-of-the-art in learning compatible features for both single and sequential multi-model upgrading.

In the following, in Sec. 2, we discuss the main literature on compatible representation learning and highlight the distinguishing features of our solution. In Sec. 3, we present in detail the problem of learning compatible representations and define new criteria and metrics for compatibility evaluation in sequential multi-model upgrading. In Sec. 4, we present our solution for learning compatible representations by exploiting feature stationarity. In Sec. 5, we evaluate our solution against state-of-the-art methods on different benchmark datasets and network architectures and demonstrate its superior performance in learning compatible representations. Finally, in Sec. 6, we perform an extensive ablation study.

Don’t hurry

Don’t anticipate complex concepts with partial and obscure sentences

Either take room to explain properly or (preferred) raise expectation

In the Introduction of the draft

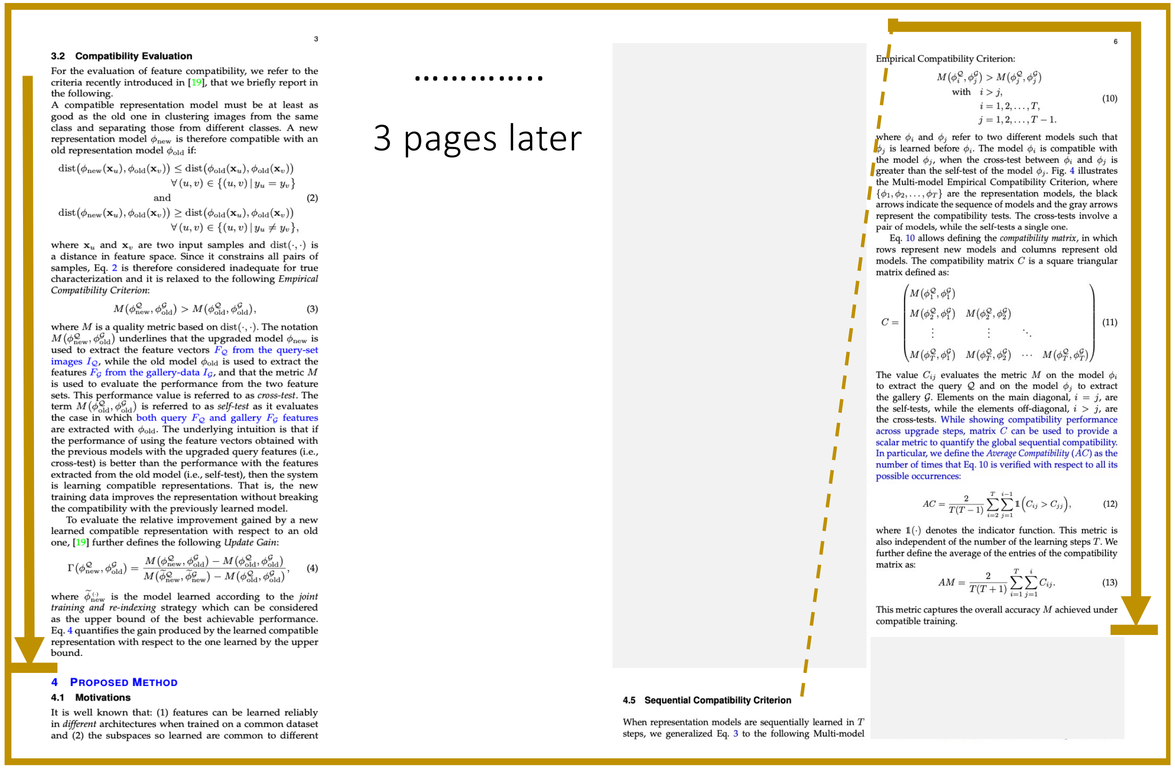
these approaches, we address compatibility by encouraging *stationarity* on the learned internal representation. Stationarity allows features’ distribution not to change under time shift so that the current learned features can be directly compared with the old ones. We argue that the stationary properties of the feature representation, *emerged* in our preliminary exploration [26], are crucial for sequential learning of feature compatibility. In particular, our training methodology learn stationarity based on two main properties of a certain family of classifiers in which the parameters are not subject to learning (i.e., fixed). The first property allows learning stationary features that exhibit strong performance in achieving compatibility; the second one allows reserving a dedicated representation space to future/unseen classes that further promote stationarity when upgrading the representation with novel classes. We extensively evaluate the compatible features learned by our training procedure on large-scale verification and identification benchmarks. We specifically evaluate the single and the sequential multi-model upgrading obtaining a large relative improvement over previous state-of-the-art. We called our method Compatible Representations via Stationarity (CoReS).

Paper body

- What may be clear to the researcher who conducted the study may not be clear to the reader
- Comment the literature in detail within a separate section where comparing your research you can enhance the interpretation of your results
- Put motivations of your work at the very beginning. Don’t let the reader wandering through the paper: let him/her understand what you present and why
- Both the reader and reviewer want to immediately understand the results of the research conducted by the author

Out of place

In the body of the draft



Put consequential subjects close each other

Don't let the reader forget what you explained before

In the final version

3 Compatibility Evaluation

We indicate with $I_{\mathcal{G}} = \{\mathbf{x}_i\}_{i=1}^N$ and $F_{\mathcal{G}} = \{\mathbf{f}_i\}_{i=1}^N$ respectively the set of images and their features in the gallery-set $\mathcal{G}$. The gallery-set $\mathcal{G}$ might be grouped into a number of classes or identities L according to a set of labels $\mathcal{Y} = \{y_i\}_{i=1}^L$. We assume that the features $F_{\mathcal{G}}$ are extracted using the representation model $\phi_{\text{old}} : \mathbb{R}^D \to \mathbb{R}^d$ that transforms each image $\mathbf{x} \in \mathbb{R}^D$ into a feature vector $\mathbf{f} \in \mathbb{R}^d$, where d and D are the dimensions of the feature and the image space, respectively. Analogously, we will refer to $I_{\mathcal{Q}}$ and $F_{\mathcal{Q}}$ respectively as the set of images and their features in the query-set $\mathcal{Q}$. The model ϕ_{old} is trained on a training-set $\mathcal{T}_{\text{old}}$ and used to perform search tasks using a distance $\text{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ to identify the closest images to the query images $I_{\mathcal{Q}}$. As novel images $\mathcal{X}$ become available, a new training-set $\mathcal{T}_{\text{new}} = \mathcal{T}_{\text{old}} \cup \mathcal{X}$ is created and exploited to learn a new model $\phi_{\text{new}} : \mathbb{R}^D \to \mathbb{R}^d$ that improves (i.e., upgrades) the ϕ_{old} model. Our goal is to design a training procedure to learn a compatible model ϕ_{new} so that any query image transformed with it can be used to perform search tasks against the gallery-set directly without re-indexing, i.e., without computing $F_{\mathcal{G}} = \{\mathbf{f} \in \mathbb{R}^d \,|\, \mathbf{f} = \phi_{\text{new}}(\mathbf{x}) \,\forall\, \mathbf{x} \in I_{\mathcal{G}}\}$.

3.1 Compatibility Criterion

In [18], a general criterion to evaluate compatibility was defined. According to this, a new and compatible representation model must be at least as good as its previous version in clustering images from the same class and separating those from different classes. A new representation model

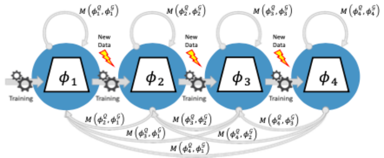


Fig. 2. Multi-model Empirical Compatibility Criterion (Eq. 5): representation models ϕ_i with $i = 1, 2, \ldots, T$ are sequentially trained. Gray arrows represent self and cross-tests (example with $T = 4$).

ϕ_{new} is therefore compatible with an old representation model ϕ_{old} if:

$$\begin{aligned} &\text{dist}\big(\phi_{\text{new}}(\mathbf{x}_u), \phi_{\text{old}}(\mathbf{x}_v)\big) \leq \text{dist}\big(\phi_{\text{old}}(\mathbf{x}_u), \phi_{\text{old}}(\mathbf{x}_v)\big) \\ &\qquad\qquad \forall\,(u,v) \in \{(u,v) \,|\, y_u = y_v\} \\ &\qquad\qquad\qquad \text{and} \\ &\text{dist}\big(\phi_{\text{new}}(\mathbf{x}_u), \phi_{\text{old}}(\mathbf{x}_v)\big) \geq \text{dist}\big(\phi_{\text{old}}(\mathbf{x}_u), \phi_{\text{old}}(\mathbf{x}_v)\big) \\ &\qquad\qquad \forall\,(u,v) \in \{(u,v) \,|\, y_u \neq y_v\}, \end{aligned} \tag{1}$$

where $\mathbf{x}_u$ and $\mathbf{x}_v$ are two input samples and $\text{dist}(\cdot,\cdot)$ is a distance in feature space. Since it constrains all pairs of samples, Eq. 1 is relaxed to the following *Empirical Compatibility Criterion*:

$$M(\phi_{\text{new}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}}) > M(\phi_{\text{old}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}}), \tag{2}$$

where M is a metric used to evaluate the performance based on $\text{dist}(\cdot,\cdot)$. The notation $M(\phi_{\text{new}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}})$ underlines that the upgraded model ϕ_{new} is used to extract feature vectors $F_{\mathcal{Q}}$ from query images $I_{\mathcal{Q}}$, while the old model ϕ_{old} is used to extract features $F_{\mathcal{G}}$ from gallery images $I_{\mathcal{G}}$. This performance value is referred to as *cross-test*. Correspondingly, $M(\phi_{\text{old}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}})$ evaluates the case in which both query and gallery features are extracted with ϕ_{old} and is referred to as *self-test*. The underlying intuition of Eq. 2 is that model ϕ_{new} is compatible with ϕ_{old} when the cross-test is greater than the self-test, i.e., by using the upgraded representation for the query-set and the old representation for the gallery-set the system improves its performance with respect to the previous condition.

To evaluate the relative improvement gained by a new learned compatible representation, the following *Update Gain* has been defined:

$$\Gamma(\phi_{\text{new}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}}) = \frac{M(\phi_{\text{new}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}}) - M(\phi_{\text{old}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}})}{M(\widetilde{\phi}_{\text{new}}^{\mathcal{Q}}, \widetilde{\phi}_{\text{new}}^{\mathcal{G}}) - M(\phi_{\text{old}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}})}, \tag{3}$$

where $M(\widetilde{\phi}_{\text{new}}^{\mathcal{Q}}, \widetilde{\phi}_{\text{new}}^{\mathcal{G}})$ stands for the best accuracy level we can achieve by re-indexing the gallery-set with the new representation [18] and can be considered as the upper bound of the best achievable performance.

3.2 Multi-step Compatibility Criterion

In real world applications, multi-step upgrading is often required, i.e., different representation models must be sequentially learned through time, in multiple upgrade steps. At each step t, the training-set is upgraded as:

$$\mathcal{T}_t = \mathcal{T}_{t-1} \cup \mathcal{X}_t \tag{4}$$

being $\mathcal{X}_t$ the new data and $\mathcal{T}_{t-1}$ the training-set at step $t-1$. In the multi-step upgrading case, we define the following *Multi-model Empirical Compatibility Criterion* as follows:

$$\begin{aligned} M(\phi_{t'}^{\mathcal{Q}}, \phi_t^{\mathcal{G}}) > M(\phi_t^{\mathcal{Q}}, \phi_t^{\mathcal{G}}) \quad \forall\, t' > t \\ \text{with } t' \in \{2, 3, \ldots, T\} \text{ and } t \in \{1, 2, \ldots, T-1\} \end{aligned} \tag{5}$$

where $\phi_{t'}$ and ϕ_t are two different models such that ϕ_t is upgraded before $\phi_{t'}$, T is the number of upgrade steps and M the metric used to evaluate the performance. Model $\phi_{t'}$ is compatible with ϕ_t when their cross-test is greater than the self-test of ϕ_t for each pair of upgrade steps. Fig. 2 illustrates the Multi-model Empirical Compatibility Criterion, where $\{\phi_1, \phi_2, \ldots, \phi_T\}$ are the representation models, black arrows indicate the model upgrades and gray arrows represent self and cross-tests.

In order to assess multi-model compatibility of Eq. 5 for a sequence of T upgrade steps, we define the following square triangular *Compatibility Matrix* C:

$$C = \begin{pmatrix} M(\phi_1^{\mathcal{Q}}, \phi_1^{\mathcal{G}}) & & & \\ M(\phi_2^{\mathcal{Q}}, \phi_1^{\mathcal{G}}) & M(\phi_2^{\mathcal{Q}}, \phi_2^{\mathcal{G}}) & & \\ \vdots & \vdots & \ddots & \\ M(\phi_T^{\mathcal{Q}}, \phi_1^{\mathcal{G}}) & M(\phi_T^{\mathcal{Q}}, \phi_2^{\mathcal{G}}) & \cdots & M(\phi_T^{\mathcal{Q}}, \phi_T^{\mathcal{G}}) \end{pmatrix} \tag{6}$$

where each entry C_{ij} is the performance value according to metric M, taking model ϕ_i for the query-set $\mathcal{Q}$ and model ϕ_j for the gallery-set $\mathcal{G}$. Entries on the main diagonal, $i = j$, represent the self-tests, while the entries off-diagonal, $i > j$, represent the cross-tests. While showing compatibility performance across multiple upgrade steps, matrix C can be used to provide a scalar metric to quantify the global multi-model compatibility in a sequence of upgrade steps. In particular, we define the *Average Multi-model Compatibility* (AC) as the number of times that Eq. 5 is verified with respect to all its possible occurrences, independently of the number of the learning steps:

$$AC = \frac{2}{T(T-1)} \sum_{i=2}^{T} \sum_{j=1}^{i-1} \mathbb{1}\big(C_{ij} > C_{jj}\big), \tag{7}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

Finally, we define the *Average Multi-model Accuracy* (AM) as the average of the entries of the Compatibility Matrix:

$$AM = \frac{2}{T(T+1)} \sum_{i=1}^{T} \sum_{j=1}^{i} C_{ij} \tag{8}$$

to provide an aggregate value of the accuracy metric M under compatible training.

Be clear and brief

Be clear and brief: use the period as a unit of measure that should not be exceeded

Avoid:

- Excessively long and complex sentences, rich in subordinate phrases prolonged with *which, and, that*…
- Abrupt transitions between topics
- Use of doubt forms with *would, could, might*….
- Parentheses that interrupt the flow of discourse
- Repeating the same thing twice or more…

In the draft

> **Class-incremental Learning (CiL).** CiL sequentially increases the number of classes to be learned by the model over time [35], [36], [37]. Although it might look similar to sequential learning of compatible features, the main focus of CiL is reducing catastrophic forgetting [38] (i.e., the tendency of a model to forget previously learned information upon learning new information). Compatible representation learning differs from CiL in two important aspects: (1) the new model is *not* required to be initialized as the old model and (2) the model has access to the *whole data* during the

Be stylish

Be stylish in writing: the choice of words and terms plays an essential role

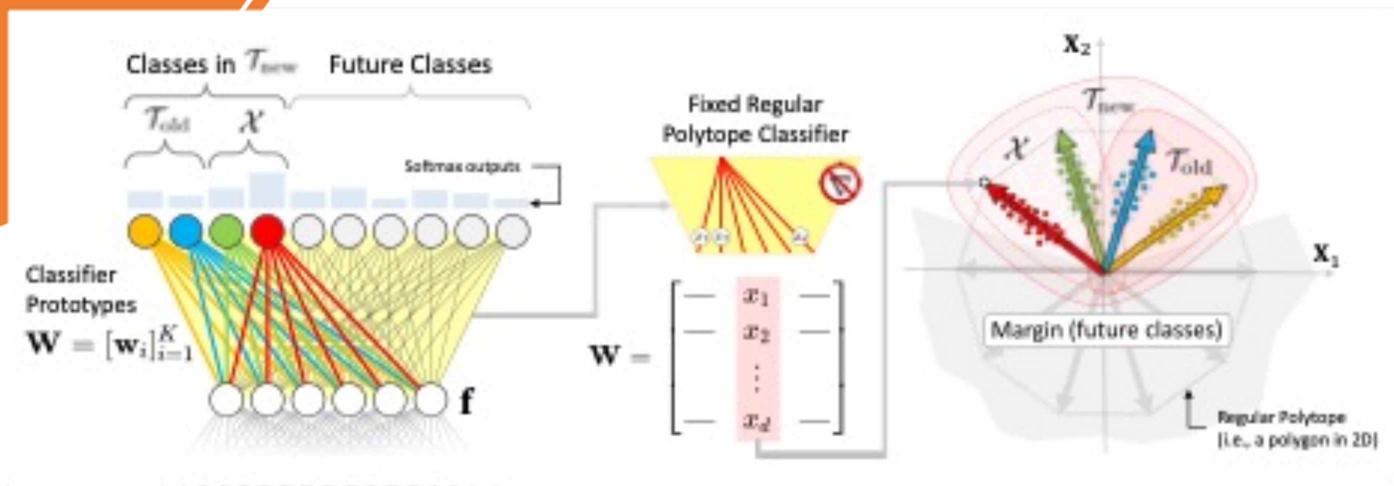
Avoid:

- Redundant and ambiguous terms
- Neologisms that are not recognized in the dictionary
- Excessive use of passive verbs
- Use of the first person
- Use of pronouns in a chain
- Use of adverbs (e.g., *absolutely, evidently, naturally*...)
- .....

Figures and captions

Average reading time of a scientific paper
~150 words per minute (3-4 minutes per page)

5



(a) (b) (c)

Fig. 4. Overview of the training procedure of CoReS based on feature stationarity showing: *(a)* fixed classifier with class prototypes $[\mathbf{w}_i]_{i=1}^{K}$ with evidence of those reserved for $\mathcal{T}_{old}$ and the upgrade classes $\mathcal{X}$ (both create $\mathcal{T}_{new}$), and those reserved for future classes; *(b)* class prototypes and their parameters $x_1, x_2, \ldots x_d$ (parameters are the coordinate vertices of the regular polytope that defines the fixed classifier); *(c)* two-dimensional representation of the feature space generated by the fixed classifier. The colored point clouds represent the learned features. Prototypes of the future classes are represented with gray arrows. The gray region is the margin imposed by the future/unseen classes. As new features are learned they are pushed out from the margin.

~10' to understand !!!!!

Figures and captions

Long captions are like subsections

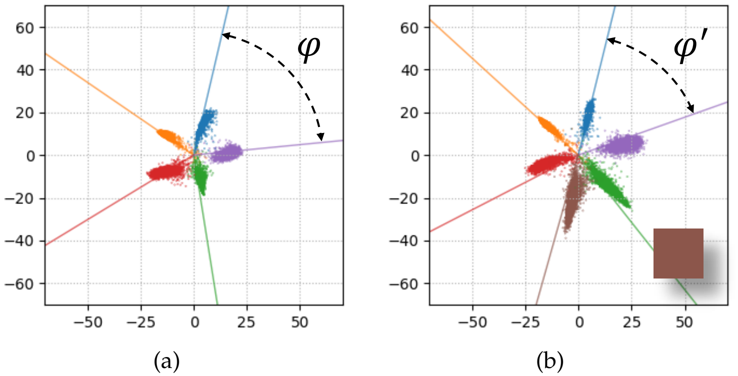


Fig. 2. Training the LeNet++ network initialized from a previously learned model (i.e., fine tuning) using the MNIST dataset. To visualize features, the output size of the last hidden layer is reduced to two. Colored cloud points are features from the test-set and colored lines represent classifier prototypes. *(a)*: Learning is performed with a training-set $\mathcal{T}_{\text{old}}$ consisting of the first five classes of the MNIST dataset. *(b)*: Learning by fine tuning $\mathcal{T}_{\text{new}} = \mathcal{T}_{\text{old}} \cup \{\text{brown-class-data}\}$. The new class determines the effect of varying the spatial configuration of the representation changing the subtended angles between features (e.g., $\varphi \neq \varphi'$).

In the draft

4

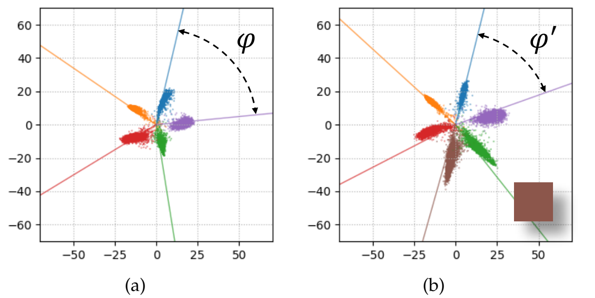


Fig. 3. Learning with incremental fine-tuning with MNIST dataset for 2D representation. Colored cloud points represent features from the test-set and gray lines represent classifier prototypes. (a) Initial configuration (5 classes); (b) Training by fine-tuning (adding the brown-class). The addition of the new class modifies the spatial configuration and angles between features.

In the final version

Be humble

After publishing the paper in IEEE TPAMI the authors developed the subject and started writing a new paper

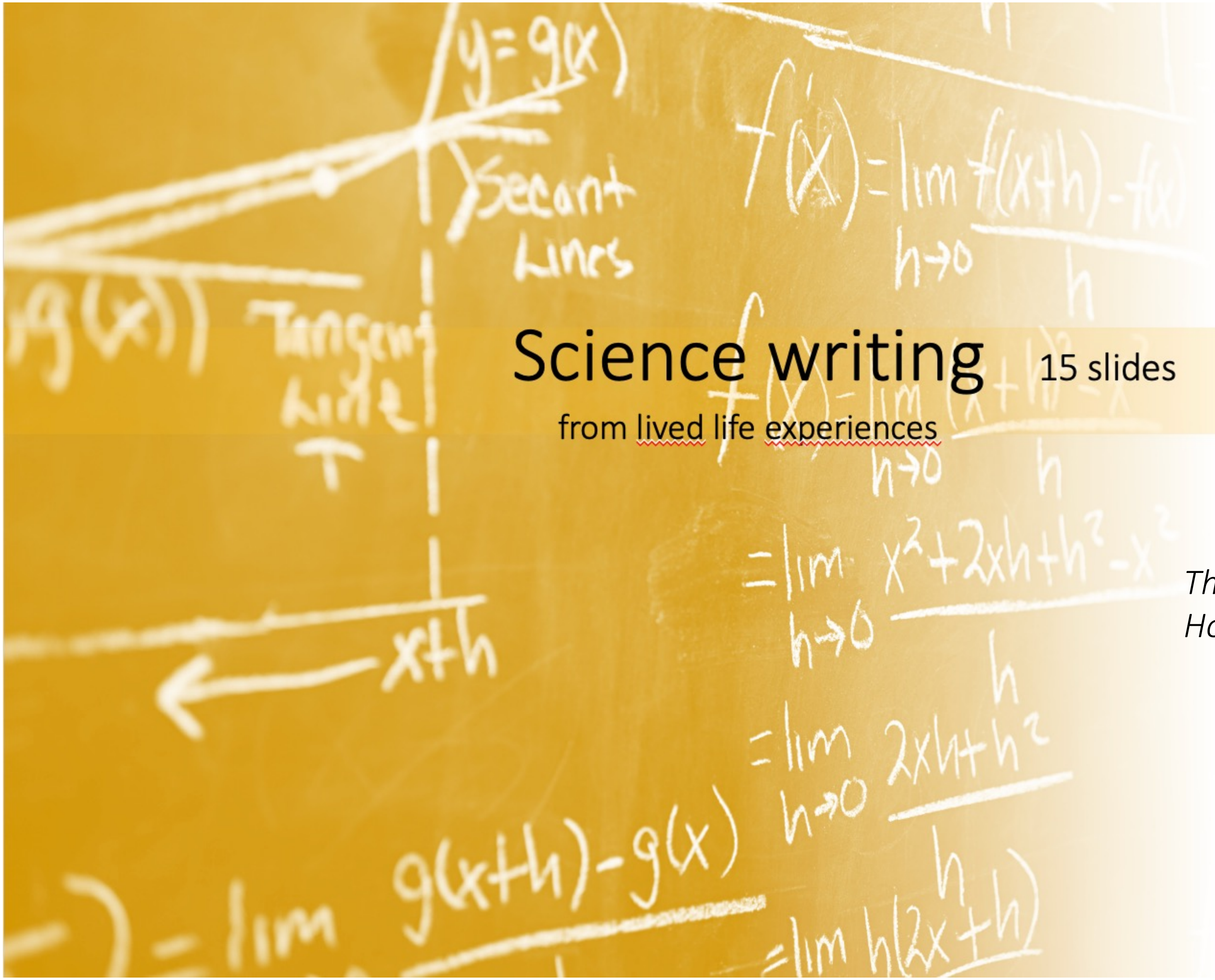
The authors were very proud of their previous results so they put references to their previous works in the new paper

In the new draft

References
[22] #5
[26] #7
[27] #2
[28] #3

!!!!!!!!!!!!!

Self-citations improve your h-index but they put the reviewer in an unfavorable disposition towards you. He/she will try to find every small defect or inconsistency in your work



Science writing

15 slides

from lived life experiences

Alberto Del Bimbo
Università di Firenze

Thank you for your attention
Hope it will help