POLITEHNICA UNIVERSITY OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 718 **from** 02-08-2021

PH.D. THESIS

Eng. Liviu-Daniel ŞTEFAN

REŢELE NEURONALE ADÂNCI PENTRU CLASIFICAREA DATELOR MULTIMEDIA

DEEP NEURAL NETWORKS FOR MULTIMEDIA CLASSIFICATION

THESIS COMMITTEE

Prof. Dr. Ing. Mihai CIUC Politehnica Univ. of Bucharest	President
Prof. Dr. Ing. Bogdan IONESCU Politehnica Univ. of Bucharest	PhD Supervisor
Conf. Dr. Ing. Ruxandra ŢAPU Politehnica Univ. of Bucharest	Referee
Prof. Dr. Ing. Henning MÜLLER Univ. of Applied Sciences Western Switzerland	Referee
SR. Dr. Ing. Adrian POPESCU CEA-LIST, France	Referee

BUCHAREST 2021

The thesis has been partly funded by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code125125.

Table of contents

Part I

Introduction and Theoretical Background

Chapter 1

Introduction

1.1 Domain of the thesis

The era of big data presents an important challenge for data analysis, processing and learning, as information become available in such a volume. In this context, images, video, audio, text, and other metadata are challenging to explore and analyze. To take advantage of such data, machine learning and deep learning approaches, in particular, are developed to analyze either independent modalities or the whole multimedia spectrum. Several of the consolidated studies results of state-of-the-art deep learning techniques show that they are capable of discovering complex patterns in massive data sets, owing to their superior representation capabilities for high-dimensional data, which, when combined with their classification capabilities, outperforms traditional descriptors and classifiers significantly. Such approaches become more and more feasible, transitioning from hypotheses to tangible systems and applications in many fields such as computer vision, economics, medical, to name a few. Nevertheless, there are still many emerging, niche or complex domains and issues that do not benefit from the deep learning endowment. The research communities are still exploring the full extent of what deep learning can achieve. In this context, this thesis studies and explores deep learning-based methods for multimedia classification with applications in consecrated and emerging topics. All this effort is directed to form a factual basis for every machine learning practitioner.

1.2 Motivation of the thesis

The number of applications that can be developed using machine learning, and particularly deep learning, is virtually endless. Nonetheless, many more applications are entirely beyond the capabilities of existing deep learning approaches. One such sector is medicine, where several automated segmentation registration methods have been

investigated and proposed for application in clinical settings. Even though medical images provide a lot of information, they are difficult to use in clinical practice due to their inherent unpredictability. Another example is the financial domain, where the data represented by commodity price, stock price, or exchange rate is unknown at $t+1$ making it extremely difficult to train deep learning approaches considering that they are notoriously data-hungry. On top of the aforementioned challenges, there is another factor that plays an important role in beating the human performance barriers with deep learning in more specialised domains. This is the lack of deep learning practitioners. One reason is this domain's inherently difficult interdisciplinary nature, requiring strong applied mathematics, programming, and domain-specific knowledge to understand, build, run, and validate approaches based on deep learning.

All these aspects motivated me to study these extremely challenging domains and collect in this thesis valuable information regarding training supervised deep neural networks alongside good practices capable of boosting networks performances, lessons learned, and useful lessons for the future.

1.3 Content of the thesis

The thesis is structured as follows: Chapter 2 covers the underlying techniques that underpin practical implementations of deep learning. Part II presents my contributions to the field of medical, computer vision, machine learning, and financial, to name a few. It starts with methods related to the medical domain involving tuberculosis types detection, multi-drug resistance prediction, and lipreading in Chapter 3, followed by methods related to surveillance involving person re-identification and person search in Chapter 4, methods related to the multimedia data involving ensemble learning in Chapter 5, methods related to financial domain involving stock trend classification in Chapter 6, and a series of common evaluation frameworks regarding interestingness prediction, violence detection and face identification in Chapter 7. The thesis concludes with Chapter 8 with some major findings and suggestions for future study, as well as an overview of my published works and my contribution to them.

Chapter 2

Deep Neural Networks

This chapter provides a timely review and survey of deep learning architectures, their applications and limitations. It aims to provide the readers with a background on various designs and the latest progress and achievements.

2.1 Convolutional Neural Networks

Structurally, CNNs contain a list of convolutional layers stacked one over the other, with each layer arranged in a 3D way, namely *width, height, depth*, similarly to an RGB image. This arrangement allows the network to recognize sophisticated shapes sequentially, from the initial pixel values to the predicted class scores. The following subsections describe such deep learning architectures in timely review and survey.

2.1.1 CNN Building Blocks

Convolution is a linear operation that measures how well two functions overlap at different locations. CNNs are usually used with images, 3D tensors with two spatial coordinate indices and one channel selection index. **Pooling** is a function expressed as a fixed shaped array that traverses all regions in the input, performing a down-sample of the input by yielding a single value at each location. **Filter hyperparameters** are composed of: (i) the dimensions of the filter, (ii) the stride which denotes the steps by which the filter traverse the input image and, (iii) the padding which denotes the operation of bordering an image with zeros.

2.1.2 VGG

The VGG architecture proposes small (3×3) convolution cores with the justification that a stack of two 3×3 filters has a 5×5 size receptive field, and a stack of three 3×3 convolutional layers has a receptive field of size 7×7.

2.1.3 GoogleNet

GoogleNet [38] consists of 22 layers with a relatively complex architecture called "inception" that aim to solve the issues imposed by the principle of depth of neural networks, namely: (i) the large number of parameters, and (ii) computing resources.

2.1.4 ResNet

ResNet [13] advents that deeper networks must be at least as efficient as equivalent shallower networks. It achieved this by learning the residual difference between the input and the mapping function.

2.2 Recurrent Neural Networks

Recurrent neural networks (RNNs)are powerfull and robust algorithms capable of remembering their input by using an internal memory, giving them the ability to be precise in predicting future sequences.

2.2.1 Long Short-Term Memory

Long short-term memory networks networks [14] represent an incremental improvement of recurrent networks that allow the extension of the memory. This improvement allows the network to learn long-term dependency.

2.3 Attention Mechanisms

Attention mechanisms [40] are essentially functions that map relevant context in the form of queries and key-value pairs to an output. Assuming there is an arbitrary query, attention mechanisms learn to attend to intermediate representations via attention pooling.

2.4 Good Practices for Deep Learning

This section investigates common methodologies and good practices for training deep learning architectures.

2.4.1 Data augmentation

One can produce additional data from the limited existing samples via various geometric transformations, color space transformations, kernel filters, mixing images and random erasing [28], to name a few.

2.4.2 Hyper-parameter tuning

The main hyperparameters are the following: (i) **Learning rate** which controls how fast a model reaches convergence, (ii) **Number of Epochs** which controls how many times the weights are updated, (iii) **Hidden Layers** which connect the input and the output layers of a neural network, Their role is to apply nonlinear transformations of the data. and (iv) **Activations Functions** which introduce non-linearity in the network.

2.4.3 Transfer learning

The motivation of transfer learning is that the network will be able to generalize faster on fewer examples if it uses the variations on rich, well-defined tasks.

2.4.4 Fine-tunning

Fine-tunning is a deep learning methodology consisting of using the weights of an existing neural network trained on a source data and initializing a new model of the same network on the target data from the same domain.

2.5 Current Limitations of Deep Learning

Every instance in deep learning is represented as a vector, which means that everything can be viewed as a point in a geometric space. Model inputs (which might include sound, text, videos, etc.) and targets are transformed into some initial input vector space and a target vector space, respectively. Each layer in a deep learning topology performs just a single geometric transformation on the data that passes through it throughout its computation. When combined, the sequence of layers results in a very complicated geometric transformation that seeks to fit the input to the target space, using a number of different transformations. It is possible to parametrize this transformation by updating the weights of the layers repeatedly, based on the network performance at a point in time. It is crucial to note that this geometric transformation must be differentiable to update parameters using an optimization algorithm for finding a local minimum. Another limitation of deep learning models is that they do not have any comprehension of the data they receive. Furthermore, when it comes to generalization, deep learning models are only competent in local generalization, which means they can only adapt to new situations that are very similar to previous data.

2.6 Conclusions

Supervised deep learning has had a significant impact in the last decade, pushing state-of-the-art systems in speech recognition, computer vision, recommender systems, language understanding and medical image analysis, to name a few many fields. While different aspects of supervised deep learning are well researched, becoming a practitioner requires a certain level of expertise to avoid overfitting the model and achieve desirable results. Furthermore, training supervised models is a laborious and time-intensive task, as the predictive capabilities of such systems depend on the quality of the data. It is well known that supervised deep learning architectures are data-hungry, and such data may have a certain likelihood of human error, impacting the quality of the predictions. Despite these shortcomings, supervised deep learning is pushing the boundaries in many industries and domains. Regarding the current limitations, the reader should note that the only concrete achievement of deep learning thus far has been the capacity to perform input-to-output mappings using a continuous geometric transform in the presence of enormous volumes of human-annotated data. While successfully doing this is transformative for virtually every field, it is still a long way from human-level performances without tackling other barriers such as reasoning and abstraction.

Part II

Personal Contributions

Chapter 3

Medical Applications

3.1 Deep Learning for Finding and Classifying Tuberculosis Types for a Targeted Treatment

3.1.1 Introduction

ImageCLEF[1] is an image retrieval and analysis assessment campaign in which algorithms are evaluated on an equal footing. One of the medical tasks at ImageCLEF 2017 was focused on tuberculosis (TB) data processing from chest CT (Computed Tomography) images. Two distinct subtasks were proposed to the participants: (1) *Multi—drug resistance* (MDR), and (ii) *Tuberculosis type* (TBT). The reader can find here [33] the main paper.

My contributions to these tasks are the following: i) exploration of fully automatic methods for segmentation of the lungs in CT volumes, ii) proposed an RGB modality that allows for applying good practices for training DNNs methods, and iii) fine-tune SOA DNNs on medical data achieving second place teamwise in the TBT prediction competition.

3.1.2 Proposed approach

We developed our approach on top of the effective GoogleNet architecture described in [38]. We chose this design due to its better resource efficiency, which enables us to expand the depth and width of the network while maintaining a consistent computational budget, allowing for efficient training in a reasonable period. In terms of volume structure modeling, a technique based on sparse structural sampling proved advantageous alongside with various training methods to address the challenges associated with a small number of training examples. These techniques include cross-modality pre-training and

[1] http://www.imageclef.org/

data augmentation with improved features. We use the supplied lung mask for both subtasks and extract the slices using a sparse structural sampling approach . Next, we extract features from each slice of the patients' CT volume using a deep CNN. Finally, we average the scores and perform training-classification on new instances using a Softmax classification.

Multi-Drug Resistant Detection

Run 1: Rank 7: In net training, we used data augmentation by sampling a 224×224 sub–image randomly cropped from the selected frame. In the test phase, given a volume, we extract all the slices. The class scores for the whole volume is then obtained by averaging the scores across the slices.

Tuberculosis Types Detection

Run 1: Rank 3: In this run, we used the RGB transformation as the input of the network finetuned on the ImageNet dataset. At test time, given a volume, we extract all the slices (samples). The class scores for the whole volume are then obtained by averaging the scores across the slices. Finally, we use a softmax classification to achieve the training–classification of new instances. This yields an accuracy of 46% on the validation set.

Run 2: Rank 8: In this run, we used a grayscale transformation by discretizing the slices into the interval from 0 to 255 via a linear transformation. We pretrained the network using the ImageNet dataset and modifying the weights of the resulted convolution layers of the ImageNet RGB model. At test time, given a volume, we extract all the slices and averaged the scores across the slices and crops therein. Finally, we use a Softmax classification to achieve the training–classification of a new volume. This yields an accuracy of 41% on the validation set.

3.1.3 Results

In the MDR subtask, the proposed approach obtained the 7th position according to the accuracy (0.5352) out of 28 participant runs. For the TBT subtask the submitted runs that obtained the 3rd and 8th position out of 23 runs submitted for this task, with a top Kappa value of 0.2329.

3.1.4 Conclusions

In this study, we present two completely automated tuberculosis categorization techniques and one automatic predictor for estimating the likelihood of TB patients developing multi-drug resistant TB. We evaluated various effective methods for training CNNs. The methods obtained an accuracy of 38.7% on the TB type dataset and 51% on the MDR dataset when using the suggested training procedures.

3.2 Deep Learning for Multi–Hypothesis Captions Prediction in Medical Images

3.2.1 Introduction

Visual interpretation and summarizing has generated increasing attention in the computer vision and natural language processing communities due to the advent of automated understanding of images and the availability of reasonable explanations. In this context, automated captioning of medical images promises to assist healthcare practitioners with important insights, while reducing their overall workflow impact. Due to its vital role, the Medical Image Captioning challenge investigated in ImageCLEF 2017 [18] advances medical image information mapping methodology in pursuit of improved methodologies for summarizing visual information from medical images. For more information, the reader is directed to the main paper [35].

My contributions consist of i) proposing a multi–hypothesis captions prediction DNN for medical images, ii) a method for codding the concepts based on one-hot encoding approach, and iii) exploration of the National Library of Medicine for building a hierarchy of concepts.

3.2.2 Proposed approach

To solve this task, we address the semantic gap between captions (text) and the image. With this in mind, we present a flexible deep CNN that accepts an arbitrary number of hypotheses as input and produces final multi-label predictions at the output. Our method takes advantage of the ResNet–152 [13] model. In this context, we pretrained the network on the ImageNet dataset and used it as the network's initialization, with a Sigmoid Cross Entropy Loss Layer in the training phase and a Sigmoid layer in the test phase, producing a probability distribution across the classes. The labels were encoded using the one-hot-encoding technique. Finally, we constructed a concept list for the test images by experimentally setting a threshold over the scores.

For all trials, we present the official metric, namely F1 score. For our approach, we pretrained the network on the ImageNet dataset and finetuned it on the target dataset by taking images of size 256×256 pixels without performing any data augmentation during training. No external resources are used to augment the data set in our approach.

3.2.3 Results

We submitted one run to the concept identification task as part of our participation. Using the protocol descrived in the previous section, we achieved an F1 score of 0.089.

3.2.4 Conclusions

This section introduces a framework for addressing the issue of multilabel image categorization. However, owing to the high number of parameters to be learnt, training a multilabel CNN is not suitable on noisy datasets with a limited number of training examples. Therefore, we demonstrate how a CNN trained on single-label image datasets, such as ImageNet, may be used to the multi-label issue to alleviate the aforementioned issue.

3.3 Deep Learning for Lip Reading: Toward Language - independent Lip Reading

3.3.1 Introduction

Lip reading is the process of interpreting spoken words from visual information, in particular visemes that are represented by lip dynamics, with many practical applications including hearing aids and helping hearing impaired individuals. In this study, we propose to exploit existing language knowledge and generalize to new languages via transfer learning by leveraging the knowledge from a source domain to a target domain. Our intuition is that such techniques will relax the hypothesis that the training data must follow certain rules such as being independent and identically distributed with the test data. Furthermore, we have generated a small scale multilingual dataset called LRM (Lip Reading Multilingual), based on the already available lip reading dataset to validate our hypotheses.

My contribution comprises of i) analyzing and implementing good practices for training deep neural network and tailored data reprocessing for visual lip reading, and ii) developed visual lip reading models for heterogeneous hardware using Graph Lowering techniques. to prove the feasibility of such systems in the context of the massive computationally intensive demand for both training and inferencing such models. To the best of our knowledge, we were the first research group to release a lip reading dataset and models for Romanian language. The reader can find here [20] the main paper.

3.3.2 Proposed approach

We explore a popular lip-reading approach, namely the D3D (DenseNet 3D) [31] followed by a set of good practices for training lip-reading models for visual speech recognition. In this regard, we discarded redundant information in order to concentrate on the discriminative region. To recognize facial landmarks, all of the images from the sequences are first passed through the MTCNN face detector [44] and aligned using the detected landmarks, and finally normalized to the overall mean and variance. The resulting mouth images are augmented using affine transformations and padded to the

Table 3.7 Classification accuracy on the multilingual LRM dataset. In bold are the best results.

Model	Invdividual subset	Acc. Top-1			
		hard		hard+easy	
		Val	Test	Val	Test
D3D	LRM	0.625	0.620	0.752	0.745
	LRW	0.835	0.830	0.836	0.847
	LRRo	0.578	**0.614**	0.885	0.892
	LRW-1000	0.463	0.418	0.535	**0.496**

same length to effectively unify to mimic variations across the multilingual datasets. These stages were established experimentally after evaluating a variety of configurations.

We validate our hypotheses using three publicly available word-level datasets, LRW [2] for English, LRW-1000 [43] for Mandarin, and LRRo [19] for Romanian. Furthermore, by combining the data subsets we create a multilingual dataset. The purpose of this dataset is to determine whether existing techniques are capable of learning the composition rules for various languages and then using that information to learn to fill in the blanks when necessary using similar language rules. For all trials, we present three measures: the top-1 accuracy, top-5 accuracy, and the Cohen's Kappa coefficient to account for the data variety.

3.3.3 Results

We provide the study results for the multilingual dataset in this section.

Multilingual Learning

On the hard subset of the LRRo, the multilingual transfer strategy outperformed the transfer learning strategy by 14.1 and 0.2 percentage points, respectively. A similar increase of 8.4 and 18.5 points was observed in the easy+hard subset. Finally, we observe an improvement of 1.1 points for the D3D on the hard+easy subset of the LRW-1000 dataset.

3.3.4 Conclusions

The results show that transfer learning improved with each of the learning protocols on both, the individual subsets and the multilingual dataset, creating a general dataset is another way to deal with unbalanced data and multilingual learning is word length-independent, making it possible for multiple languages to be utilized.

Chapter 4

Video Surveillance

4.1 Joint Person Detection and Re-identification

4.1.1 Introduction

Person re-identification addresses the problem of searching specific persons across spatially non-overlapping cameras, by estimating visual similarities between different probe-gallery pairs.

My contribution consists of proposing an end-to-end framework for person re-identification that perform both person detection and classification, in contrast to general approaches that use an oversimplified protocol which assume that the spatial coordinates of each person from the gallery are given and perfectly aligned. The reader can find here [36] the main paper.

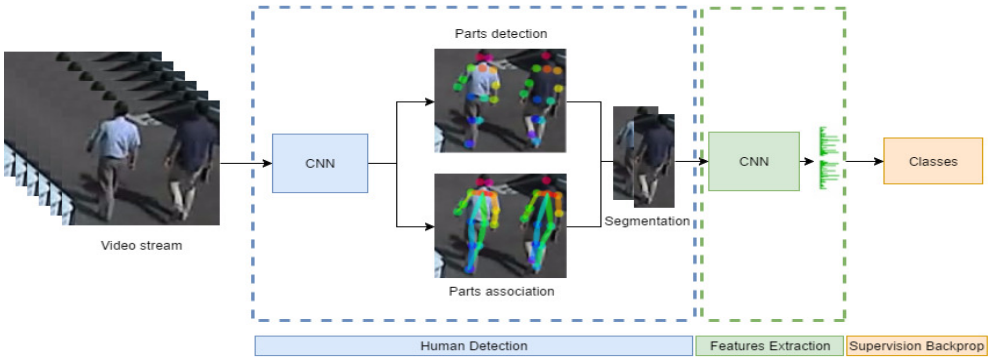


Fig. 4.1 The proposed end-to-end very deep network: (1) human detection, (2) segmentation, (3) features extraction, (4) softmax classification [36].

4.1.2 Proposed approach

The proposed system comprises of four processing stages. as depicted in Figure 4.1. In the first step, we recognize humans in video streams by predicting vector fields that directly disclose the relationship between anatomical components in an image. Our model was trained using the annotation in the MSCOCO dataset.The second step is to segment the human body components. A color image is fed into the human detection CNN to jointly predict confidence maps for body part detection and vector fields. For

each individual in the picture, the outputs are the 2D positions of anatomical key-points. We combine them into full-body postures for all of the persons in the shot and create the maximum bounding box for each of them. In the third step, we perform the feature extraction, where body images are converted into meaningful content descriptors.The final stage is feature extraction, which converts full body images to meaningful content descriptions. This was accomplished by utilizing the VGG network described in [29]. Finally, we use a softmax classification to achieve the training–classification of new instances.

Evaluation

For evaluating our system, we used SCOUTER augmented the dataset using a 10 crops strategy.To assess retrieval performance, we use a global measure of performance, namely the MAP.

4.1.3 Results

We compare our adapted network architecture with the state-of-the-art, in this section.Given the specificity of the task, i.e., automated video surveillance, we tested two approaches: i) training from scratch; ii) pre-training last layer. Our first approach, training from scratch, achieves a MAP of 60.56% whereas our second one, pre-training last layer + fine-tuning, achieves a MAP of 66.86% outperforming the other existing approaches with 5 respectively 11 percents.

4.1.4 Conclusion

In this section, we tackled the challenge of person re-identification by developing a trainable end-to-end deep neural network capable of accurately re-identifying persons in multiple-stream video from a variety of sources (indoor and outdoor).

4.2 Person Search

4.2.1 Introduction

Given a query image, person search attempts to locate the specified person in a gallery of images using the query image. It is a generalization of the traditional person re-identification classification problem [46, 17], which is predicated on two assumptions, namely: i) the spatial coordinates of each person in the gallery are provided, and (ii) the spatial coordinates are properly matched. However, these two assumptions do not hold true in actuality. In this section, we take advantage of the deep learning advancements, and propose an end-to-end person search framework that integrates multiple DNN architectures.

The contribution beyond state of the art can be summarized with the following: (i) we integrate attention mechanisms in the stem CNN to train the network to attend to representative parts in pedestrain patches. This allow the network to focus on discriminative regions, e.g., face and body parts, or on different accessories such as glasses, bags, etc.; (ii) we perform spatial transformations to increase the robustness of the system to spatial variances. The reader can find here [32] the main paper.

4.2.2 Proposed approach

In this part, we offer a thorough description of the proposed person search network. In addition, to increase the algorithm's predictive capacity, we research, test, and incorporate attention layers into the design. The methodology for conducting a unified person search consists of three parts: (i) global features, which are low-level features extracted from the entire input image, (ii) region proposals for converting low-level features into pedestrian proposals, and (iii) local features, which are discriminative features corresponding to the image's identities. We developed and evaluated three common deep neural networks (DNNs) from the literature as the backbone of our architecture, namely GoogleNet [38], ResNet50 [13], and DenseNet121 [16].

Evaluation

We evaluate our proposed method on three large-scale end-to-end person detection and re-identification benchmarks, namely: PRW [47], CUHK03 [25], and CUHK-SYSU [42]. For each data set, we adopted the original evaluation protocol that the data set provides, namely the mean average precision (mAP). Finally, the results are obtained in a single-query setting, without re-ranking.

Table 4.3 Effectiveness of the person search framework expressed in mAP (IoU $>$ 0.75).

Method	**PRW**	**CUHK03**	**CUHK-SYSU**
GoogleNet	0.263	0.668	0.685
ResNet50	0.327	0.714	0.753
DenseNet121	0.335	0.703	0.778
$GoogleNet_{att}$	**0.281**	**0.694**	**0.692**
$ResNet50_{att}$	**0.347**	**0.721**	**0.775**
$DenseNet121_{att}$	**0.358**	**0.717**	**0.783**

4.2.3 Results

We compare our proposed person search framework (with or without using attention mechanisms) using three consecrated deep networks, namely GoogleNet, ResNet50, and DenseNet121 as the backbone of the framework to evidentiate their utility in person search. The results are summarized in Table 4.3. DenseNet121 surpasses the other two networks on two out of three data sets. The tendency holds for variations with attention mechanisms,which consistently outperform the baseline variations.

4.2.4 Conclusions

Extensive experiments show that the attention mechanism consistently improves the overall performance of the system, achieving a mAP score of 0.358, 0.721, and 0.783 on the PRW, CUHK03, and CUHK-SYSU data sets, respectively.

Chapter 5

Ensemble Learning

5.1 Deep Learning for Ensemble Systems

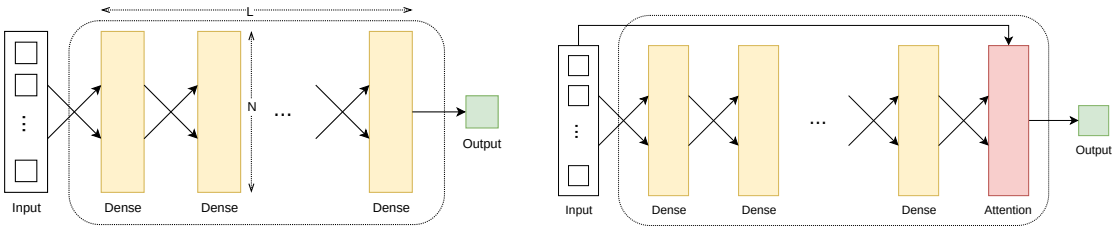
5.1.1 Introduction

An ensemble is defined as a collection of independent classifiers (inducers), each with its own area of competence, and a learning or combination algorithm that generates a new output based on the separate outputs of inducers in order to predict future incoming instances. It presents a strategy for late fusion based on a set of systems. The objective of *ensembling* is to produce a powerful learner based on the expertise of the numerous classifiers it contains. In this section, we present a deep ensembling architecture, which is a deep learning-based technique for discovering patterns and correlations between the responses of individual classifiers.

Concretely, contributions consists of: i) proposing a novel network builder of end-to-end deep architectures, The architectures are built in a progressive order, simplest models first while varying the numbers of dense layers, the numbers of neurons for each dense layer, and including or excluding batch normalization layers, ii) developing deep neural networks tailored for ensemble learning, and iii) validating the systems on a large collection of common evaluation frameworks under various formulations, from binary classification/regression to multi-label classification. At publishing time, these methods outperformed literature state-of-the-art with a great margin on all the data sets. The reader can access the full papers here [4, 3, 34].

5.1.2 Proposed approach

The proposed method follows the hypothesis that by aggregating using a deep neural network architecture, we can more efficiently describe the bias learnt by each system and the correlations across the biases, all while lowering the variance, allowing for robust retrieval. To do this, we utilize straightforward yet efficient deep neural network designs based on dense and attention ensembles.



(a) Dense Architecture.

(b) Dense Architecture with Attention Mechanisms.

Fig. 5.2 Proposed deep ensembling architectures: variable number of dense layers (L), variable number of neurons per dense layer (N), and variable number of filters for the convolutional layer (F) [34].

Dense Architecture

Given that dense architectures are universal approximators able to learn any function, we stack such dense layers to create a first baseline ensemble network. To begin, we define a set of criteria for constructing network topologies, which include the following: i) altering the number of fully-connected layers, (ii) varying the number of neurons in each fully-connected layer, and (iii) including or removing batch normalization layers. The diagram of the implemented dense architecture is presented in Figure 5.2a.

Dense Architecture with Attention

To increase the accuracy of the ensemble architecture, we augment our baseline ensemble rules with soft attention mechanisms. In this regard, feature maps are multiplied by soft maps with values between 0 and 1. A diagram of the proposed implementation is presented in Figure 5.2b.

5.1.3 Predicting visual interestingness

The MediaEval 2017 Predicting Media Interestingness data set [8] provides data for two scenarios: (i) prediction of image visual interestingness (*INT2017.Image*), and (ii) prediction of video visual interestingness (*INT2017.Video*).We have considered all the systems participating in the benchmarking campaign, namely 33 systems for the INT2017.Image, and 42 systems for the INT2017.Video prediction tasks.

Evaluation

The evaluation is carried out using the following test data split scenarios: (i) 75% training and 25% testing (RSKF75), and (ii) 50% training and 50% testing (RSKF50). Split samples are randomized, and this action is performed multiple times to obtain a thorough coverage. In the end 100 partitions are generated. The metrics are computed as average values over these partitions.

Results

This section presents the results of the best-performing architectures. The achieved results are summarized in Table 5.1. For the INT2017.Image data, the best performing dense architecture uses 10 dense layers and 1,000 neurons per layer, without batch normalization, achieving a mAP@10 of 0.3355 for the RSKF75 split and of 0.2316 for RSKF50. The addition of attention layer further increased the results, the best performing architecture with the RSKF50 split achieving a mAP@10 of 0.2399. For the INT2017.Video data, the best performance with a dense architecture is achieved using 25 layers and 2,000 neurons per layer, with batch normalization, yielding a mAP@10 of 0.2677 and 0.1562, for RSKF75 and RSKF50, respectively.

5.1.4 Predicting violent scenes

The MediaEval 2015 Affective Impact of Movies (*VSD2015.Video*) data set [30] is composed of 31 full movies, 86 YouTube videos and 10,900 short clips extracted from 199 movies of various genres (up to 96 hours). We have considered all the 48 systems trained in the competition.

Evaluation

For assessing performance, we use the official metrics released by the authors of the data (on which the inducers were optimized), name the Mean Average Precision. The metrics are computed using the standard treceval software tool.

Results

The achieved results are summarized in Table 5.1. For the VSD2015.Video data, an architecture with 5 dense layers and 500 neurons per layer achieved the highest performance for the dense architecture, with a mAP of 0.6341 for the RSKF75 split and of 0.6192 for the RSKF50. The best results for the RSKF75 split are obtained with an attention architecture, mAP of 0.6486.

5.1.5 Predicting emotional impact of movies

The MediaEval 2018 Emotional Impact of Movies [7] is a data set for automatic recognition of emotion in videos, in terms of valence, arousal, and fear. We have considered all the systems participating in the benchmarking campaign, namely 30 systems for the valence and arousal prediction task, and 18 systems for the fear detection task.

Evaluation

To evaluate the performance of the proposed models, we have adopted the official metrics released by the authors of the data, namely: (i) for the MediaEval 2018 Emotional Impact

Table 5.1 Results for the INT2017Image and INT2017Video with mAP@10 metric, VSD2015 with mAP metric, Caption with F1 metric, Arousal-Valence, with MSE and PCC metrics and Fear, with IoU metric.

System	Split	INT2017 Image (mAP@10)	INT2017 Video (mAP@10)	VSD2015 Video (mAP)	Caption (F1)	Arousal (MSE)	Arousal (PCC)	Valence (MSE)	Valence (F1)	Fear (IoU)
SOA	orig	0.1560	0.0930	0.3030	0.2823	0.1334	0.3358	0.0837	0.3047	0.1575
Best-LF	KF75	0.1674	0.1129	0.3920	0.2846	0.1282	0.3911	0.0769	0.3972	0.1621
DF-Dense	KF75	0.2316	0.1563	0.6192	0.3462	0.0549	0.8315	0.0626	0.8101	0.2129
	KF50	0.3355	0.2677	0.6341	**0.3740**	0.0571	0.8018	0.0640	0.7876	0.1938
DF-Attn	KF75	**0.2399**	0.1668	0.6228	0.3522	0.0548	**0.8339**	**0.0626**	**0.8107**	**0.2140**
	KF50	0.3389	**0.2750**	**0.6486**	0.3659	**0.0568**	0.8036	0.0640	0.7888	0.1913

of Movies data set, we use the Mean Square Error (MSE) and the Pearson's Correlation Coefficient (PCC) for the valence and arousal prediction task, with MSE being the primary metric, and Intersection over Union (IoU) of time intervals, for the fear detection task.

Results

The achieved results are summarized in Table 5.1 . For the Arousal-Valence data, the best performing DF-Dense architecture managed to improve those results, reaching to 0.0549 and 0.0626. The DF-Attn network further improved these results. For the Fear data, the best performing DF-Dense configuration achieves a score of 0.2129, increasing the state-of-the-art result by 35.17%. Furthermore, both the DF-Attn configuration increases the score with a maximum IoU of 0.2242, representing a approx. 42% increase over the original performance.

5.1.6 Concept detection

The ImageCLEFmed 2019 Concept Detection (denoted *Caption*) is an automatic multi-label classification image captioning and scene understanding data set.We have considered all the systems participating in the benchmarking campaign, namely 58 systems.

Evaluation

To evaluate the performance of the proposed models, we have adopted the official metrics released by the authors of the data, namely, the F1-scores computed per image and averaged across all test images. The metrics are computed as average values over all the partitions.

Results

The achieved results are summarized in Table 5.1. The best performing DF-Dense configuration is represented by a 5 layer network, with 500 neurons per layer and no batch normalization. This configuration achieves an F1-score of 0.3740, increasing the state-of-the-art results by 32.48%.

5.1.7 Ablation Study

Table 5.6 Analysis of the results in correlation with the inducer performance and diversity. We present the scores recorded by the best and worst inducers, the best results for both RSKF75 and RSKF50 splits, the diversity measure ($\bar{r}$) and the percentage increase over the best performing inducer (δ_{75} and δ_{50}).

Run	INT2017.Image	INT2017.Video	VSD2015.Video
Best Ind.	0.1385	0.0827	0.296
Worst Ind.	0.0126	0.0396	0.0419
$\bar{r}$	0.225	0.1017	0.1997
RSKF75	0.3436	0.2799	0.6486
RSKF50	0.2399	0.1692	0.6281
δ_{75}	148.08	238.45	119.12
δ_{50}	73.21	104.59	112.19

An important analysis is to study the influence of the system diversity on the results. To do so, we employ the Pearson's correlation coefficient. The methodology is described in the extended version of the thesis. Given the x_i input vector for a sample i of size k, representing the output scores of each classifier, $\{f_0, f_1, ..., f_{k-1}\}$ (see equation 5.4), we choose to decorate each element of this vector with output scores and correlation scores from the most similar systems with respect to output. Therefore, given the matrix Y (see equation 5.5) where each of the y_i vectors represents the scores given by the classifier f_i for all the samples, we calculate the similarity between each classifier via the PCC. The descending ordered vector of correlations for each y_i vector would correspond to R_i (see equation 5.6), where $R_{0,i}$ is the correlation coefficient of the most similar system. $R_{1,i}$ represents the correlation coefficient of the second most similar system and so on. The final decorated version of the input is represented in equation 5.7, where, for each sample i, the pairs $(c_{0,j,i}, r_{0,j,i})$ represent the output score (c) and Pearson's correlation score (r) for the most similar system with any given system score s_j. the pairs $(c_{1,j,i}, s_{1,j,i})$ represent the second most similar system, and so on. The decorated vector xd_i size is $3k \times 3$.

$$x_i = \begin{bmatrix} s_{0,i} & s_{1,i} & ... & s_{k-1,i} \end{bmatrix} \tag{5.4}$$

$$Y = \begin{bmatrix} y_0 & y_1 & ... & y_{k-1} \end{bmatrix} \tag{5.5}$$

$$R_i = \begin{bmatrix} R_{0,i} & R_{1,i} & ... & R_{k-2,i} \end{bmatrix} \tag{5.6}$$

$$xd_i = \begin{bmatrix} r_{3,0,i} & c_{0,0,i} & r_{0,0,i} & & r_{3,k-1,i} & c_{0,k-1,i} & r_{0,k-1,i} \\ c_{3,0,i} & s_{0,i} & c_{1,0,i} & ... & c_{3,k-1,i} & s_{k-1,i} & c_{1,k-1,i} \\ r_{2,0,i} & c_{2,0,i} & r_{1,0,i} & & r_{2,k-1,i} & c_{2,k-1,i} & r_{1,k-1,i} \end{bmatrix} \tag{5.7}$$

By taking the correlation coefficient vector R_i corresponding to each classifier f_i (see equation 5.6), we can calculate the average value of this vector $\bar{R}_i$. In the final step,

the overall correlation for all the inducer systems, $\bar{r}$, is calculated as the average of all $\bar{R}_i$ values and will be used as a measure of diversity for the inducers. Just like Pearson's correlation coefficient, lower values of $\bar{r}$ indicate a lower correlation between systems and, therefore, more diversity. Table 5.6 presents the results of this analysis. The results show that, for the RSKF75 split, the highest growth (δ_{75}) is recorded by the INT2017.Video systems, which increase the best inducer's result by 238.75%. This is interesting as the inducers for the INT2017.Video date are the most diverse set out of all three tasks, $\bar{r} = 0.1017$. However, this observation is not consistent with the RSKF50 split, where the second most diverse inducers, VSD2015.Video data, $\bar{r} = 0.1997$, show the highest increase, $\delta_{50} = 112.19\%$, while the INT2017.Video inducers come in second place, with $\delta_{50} = 104.59\%$. This could be a confirmation that the RSKF50 split may provide too few samples for the tasks.

5.1.8 Conclusions

We presented a method for deep ensembling that utilizes architectures composed of dense and attention layers. The primary advantage of the proposed architecture is its ability to find connections between inducer systems automatically. The results demonstrated a significant improvement over inducer systems and state-of-the-art approaches, with many instances exceeding by at least a factor of two the mAP performance.

Chapter 6

Fintech

6.1 Deep Learning for Financial Time Series Prediction

6.1.1 Introduction

Predicting stock markets is of great interest in financial economics with several empirical studies suggesting that stock markets are predictable to some extent. This section proposes to study the potential of deep representations learning as a mechanism for stock return prediction that targets both synthetic data generation and trend prediction[1]. The main papers can be accessed here [11, 10].

6.1.2 Proposed approach

Given an entire stock universe U, we propose to analyse each stock and search a prediction function f to predict each stock return at time $t+1, r_t+1$. given the representation

[1]This work is the result of research funded by Hana TI.

f_t, which can be either a linear or nonlinear transformation of the raw data R_t extracted at time t. In our research, we use the past returns of stocks. Given M stocks in U, and g lagged returns, R is represented as $R_t = \left[r_{1,t}, ...r_{1,t-g+1}, ..., r_{M,t}, ..., r_{M,t-g+1}\right]^T$.

Synthetic data generation

We synthesize time-series with quantitative properties comparable to that of the original financial time-series by targeting the statistical properties such as distributional, dependence, pathwise, cross-sectional properties within similar ranges with the original data set. The entire process is described in [11, 10]. Our GAN model synthesises an entire window in contrast to classical approaches that synthesises the (n+1)th sample, based on the preceding n samples. Therefore, we produce a fixed-length 1d vector of log return of Close prices. We assessed the quality of the models by randomly selecting a batch of both real and synthetic data for each epoch and analyze their distribution. We further explored two more properties, namely, central moments and autocorrelation.

Unsupervised statistical clustering for industry classification

We have selected an unsupervised learning algorithm in the detriment to the industry classifications as GICS[2], NAICS[3], Finviz[4], etc., (which are independent of the pricing data) because the synthetic data (GAN based data) is unknown to the aforementioned classification schemes, therefore, we are required to use an automatic algorithm to cluster the real and synthetic stock universe. In this context, We have applied the KMeans clustering algorithm to cluster the entire universe of stock according to how close the normalized returns are to the cross-sectional means of the parent clusters as in [21]. Let N be the number of observations, d the trading days, and R_{is} are the daily stock returns, where $i = 1, \ldots, N$, and $s = 1, \ldots, d$, we cluster the normalized returns R_{is}, where $\hat{R}_{is} = \frac{R_{is}}{\sigma_i u_i}$, and $u_i = \frac{\sigma_i}{v}$. For all $u_i < 1, u_(i) \equiv 1$, and Median$(\cdot)$ and MAD$(\cdot)$ are cross-sectional. The standard deviation is computed with a loopback of 100 days, and the clusterization is performed with a loopback of 1000 days, and a stride of 30.

Statistical labeling

Let P_{is} be the time series of stock prices, where $i = 1, \ldots, N$ labels the stocks, and $s = 1, \ldots, d$ labels the trading dates, a time series from day s will be assigned with a corresponding label, denoted L_{is}, according to the value of P_{is} compared to the median of the cluster it belongs to. If P_{is} is greater or equal to the median, then $L_{is} = 1$, otherwise $L_{is} = 0$.

[2]https://www.msci.com/gics
[3]https://www.census.gov/naics/
[4]https://finviz.com/

Features extraction methods

Let w be the rolling window size, i.e., the number of consecutive observation per rolling window, T the sample size, and suppose the number of increments between successive rolling windows is 1 period, the time series is partitioned into $N = T - w + 1$ subsamples. For each day d, we use a rolling window size, w. The first rolling window contains observations for period 1 through w, the second rolling window contains observations for period 2 through $w + 1$, and so on. This feature allows us to build a baseline by estimating the prediction model using each rolling window subsamples of returns.

Denoising and dimensionality reduction

Because of the huge number of immediate market movements and trade noise, financial data has a complicated structure of irregularities and roughness. The noise in financial data generally shows strong tailedness, which means that the underlying time series data has a lot of sharp breaks every once in a while. Ignoring these anomalies might lead to erroneous data mining and statistical modeling results. As a result, In order to unveil more meaningful representations, we propose to denoise and reduce the dimensionality of the data using a stacked autoencoder structure. We set the hidden layer size to 16 and the depth of the SAE to 4.

Prediction

Three variants of DNN have been implemented and tested, namely a multi-layer perceptrons (MLP), a time-series 1-dimensional stacked separable convolution neural network based on ResNet, and finally a bidirectional Long Short Term Memory (BiLSTM) network.

Evaluation

We run our simulations over 17 years, using a two-part split protocol i) train/test on real data, ii) train on real and augmented data and test on real data, using a split of 7 years of data for training and 1 year of for testing.

Table 6.1 Prediction results expressed in accuracy with MLP for 50 epochs.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	49.38%	48.92%	50.04%	50.01%	47.64%	49.32%	50.04%	49.93%	49.41%
2	49.34%	49.61%	50.04%	50.03%	49.41%	49.65%	49.45%	49.75%	49.66%
3	49.81%	50.04%	50.03%	49.89%	49.84%	49.91%	49.58%	50.00%	49.89%
...	...	...	...	...	...	...	....	...	...
50	50.25%	50.50%	50.28%	50.63%	50.31%	50.16%	50.51%	50.78%	50.43%

Table 6.2 Prediction results expressed in accuracy with ResNet1D for 50 epochs.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	49.21%	49.43%	50.04%	50.38%	49.88%	50.20%	50.02%	50.11%	49.91%
2	49.19%	49.21%	50.01%	50.53%	50.15%	50.12%	50.02%	50.09%	49.92%
3	49.37%	49.48%	50.06%	50.51%	50.13%	50.15%	50.01%	50.19%	49.99%
...	...	...	...	...	...	...	....	...	...
50	51.84%	50.44%	50.56%	51.34%	51.14%	51.05%	51.10%	51.93%	51.18%

Table 6.3 Prediction results expressed in accuracy with LSTM for 50 epochs.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	49.91%	50.14%	50.44%	50.15%	50.01%	49.74%	50.35%	49.44%	50.02%
2	49.64%	50.12%	50.41%	50.17%	50.04%	49.11%	50.31%	49.91%	49.96 %
3	50.05%	50.14%	50.39%	50.11%	49.85%	49.89%	50.31%	50.11%	50.11 %
...	...	...	...	...	...	...	....	...	...
50	51.16%	51.55%	51.81%	51.52%	52.02%	51.38%	51.62%	52.04%	51.64%

Table 6.4 Prediction results expressed in accuracy with LSTM train for 50 epochs on synth + real data and tested on real data.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	50.21%	49.51%	50.57%	51.01%	50.45%	50.14%	48.53%	51.56%	50.25%
2	50.19%	50.01%	50.61%	51.13%	50.45%	50.26%	48.24%	51.43%	50.29%
3	50.20%	49.95%	50.04%	50.47%	50.17%	50.11%	51.19%	50.49%	50.33%
...	...	...	...	...	...	...	....	...	...
50	51.51%	52.82%	52.81%	52.42%	52.85%	51.86%	52.49%	52.91%	52.46%

6.1.3 Results

We have used [23] as an external baseline. The manuscript tests 1.000.000 machine learning models for directional forecasting using data from 1993 to 2008 to predict SPY ETF direction and achieved an accuracy of 51.5%.

Train on real data–test on real data

Training with MLP, ResNet1D and BiLSTM we achieved an overall accuracy of 50.43%, 51.64% and 52.46%, respectively after 50 epochs.

Train on real + synthetic data–test on real data

We have retrained the BiLSTM network using a mix of real and synthetic data and tested it strictly on real data achieving an an overall accuracy of 52.46% after 50 epochs. Using the synthetic data, we increased the overall accuracy from 51.64% to up to 52.46%. Furthermore, the results are significant, reporting the results to the external baseline.

6.1.4 Conclusions

This section presents a DNN framework for financial time series prediction of performing and underperforming stocks. Three topologies, namely, a multi-layer perceptron, a 1D variation of ResNet and an LSTM, were tested under two settings, trained and tested on real data and trained on a mix of real and synthetic data and tested on real data. The results show that synthetic data have a positive impact, improving the overall accuracy

from 51.64% to up to 52.46%. These results are significant, reporting the results to the literature baselines.

Chapter 7

Datasets and Evaluation

This chapter presents my contribution to the analysis, generation or publicly releasing of a collection of datasets, namely: (i) *Interestingness*10*k* [6], for interestingness prediction in multimedia data ; (ii) *VSD*96 [5], for violent scenes detection, and finally, (iii) *FaVCI2D* [26] for face verification with challenging imposters and diversified demographics. Furthermore, it provides insights, observations and recommendations concerning the development of systems to predict the relevant representations for each of the datasets.

7.1 Interestingness prediction

Dataset description

The Interestingness10k [6] dataset is a freely available dataset and a testbed for predicting the interestingness of images and videos. This dataset was evaluated and validated during the MediaEval Predicting Media Interestingness tasks in 2016 and 2017. My contributions to this data set comprises of: i) analyzing the employed techniques and their capabilities on the *Interestingness10k* data set, that allow for general trends to be deduced with respect to the best performing systems ii), providing insights about the capabilities of the newest deep neural networks, by analyzed the performance of state-of-the-art architectures on the interestingness10k data set, iii) providing a series of recommendations in respect to systems performances, iv) analyze and visualize how algorithms interpret interestingness, v) developing an ensemble method based on the runs submited to the MediaEval Predicting Media Interestingness and, vi) taking part in the annotation process.

7.1.1 Problem formulation

According to the problem formulation, we identified 4 main approaches: more than 52% of the approaches use *classification*, ranking for 31% (it is worth noting that ranking approaches achieve overall better results in all the cases compared to the classification (Mann-Whitney-U $p < 0.005$)), regression with 15% and hybrid with 2% of the total number of systems.

7.1.2 Methods analysis

According to the classification problem, we identified the following approaches: Support Vector Machines accounting for 30%, Deep Neural Networks for 28%, Ranking for 13%, Regression for 12%, Hybrid for 6%, Neural Networks for 6%, Distance-based for 4% and Ensemble learning and Statistical, each one under 1%. Looking at their overall performances, DNNs, NN and hybrid approaches were the best performers.

7.1.3 Overall method performance analysis

For the image data, approaches based on DNNs and shallow NN stand out, with average mAP scores of 0.2460 and 0.2405, respectively. On the other hand, for the video data, hybrid approaches and SVM-based approaches stand out as the best performers, with average mAP scores of 0.1867 and 0.1822, respectively.

7.1.4 State-of-the-art deep neural networks

To account for current state-of-the-art deep neural network capabilities, we evaluate the performance of three recent image and video classification architectures. We selected for the image data the ResNeXt-101-32x48d , PNASNet-5 and ResNet-50 architectures, augmented with best practices as presented in [39], and for the video data, the GSM-InceptionV3 En3 , IR-CSN-152 , and R(2+1)-18 architectures. The analysis of the results shows that these deep neural networks do not achieve the best results.

Table 7.3 Performance of state-of-the-art deep neural network architectures trained on the Interestingness10k data.

	Method	2016 (mAP)	2017 (mAP@10)
Image	bestME	0.2336	0.1385
	bestSoA	0.2485	0.1560
	FixResNet50 [39]	0.1906	0.1099
	FixPNASNet-5 [39]	0.1981	0.1233
	FixResNeXt-101-32x48d [39]	0.2273	0.1410
Video	bestME	0.1815	0.0827
	bestSoA	0.1815	0.0930
	IR-CSN-152 [12]	0.1577	0.0629
	R(2+1)-18 [12]	0.1579	0.0644
	GSM-InceptionV3-En3 [37]	0.1738	0.0821

To understand how deep learning algorithms interpret the visual samples and thus how they attempt to predict interestingness, we computed the Grad-CAM and Guided Backpropagation Grad-CAM maps. Some relevant examples are presented in Figure 7.2. Results show that in many cases, the model focuses on the main subject, but predominantly more on elements adjacent to it, showing an inclination for detecting the context

that surrounds the main subject. This is also true for human subjects, as the Grad-CAM analysis shows network activation on human faces, but also many times around the face. We theorize that this concentration of useful features on and around faces may represent a positive influence on the final results, as faces convey emotions.

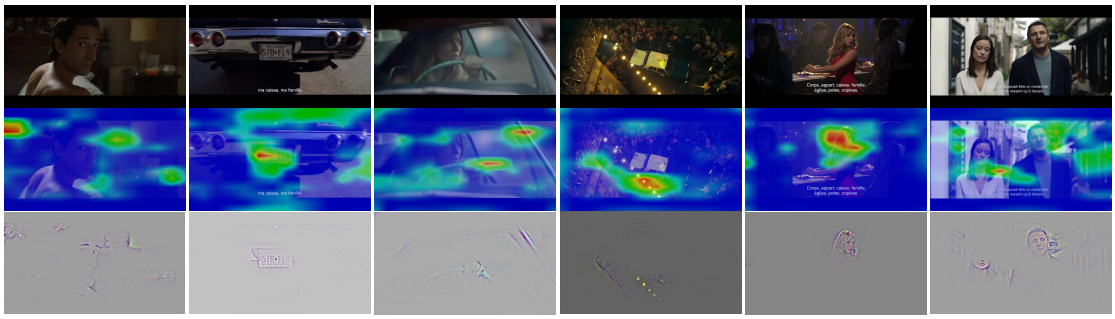


Fig. 7.2 Grad-CAM analysis of the network interpretation of visual information in the case of images predicted as interesting. The top row presents the original samples, the middle row presents the Grad-CAM output image describing "class-discriminative regions", and the bottom row presents the Guided Backpropagation Grad-CAM describing the features that most contributed to the class decision [6].

7.1.5 Proposed MLP architecture

We provide a straightforward, yet efficient, late fusion technique based on a deep MLP architecture. Our method is driven by the dense layer's ability to uncover patterns and connections among individual system outputs. We propose to model the bias learnt by each inducer as well as the correlations between the biases in order to conduct retrieval consistently and improve the aggregated system's overall performance. After testing

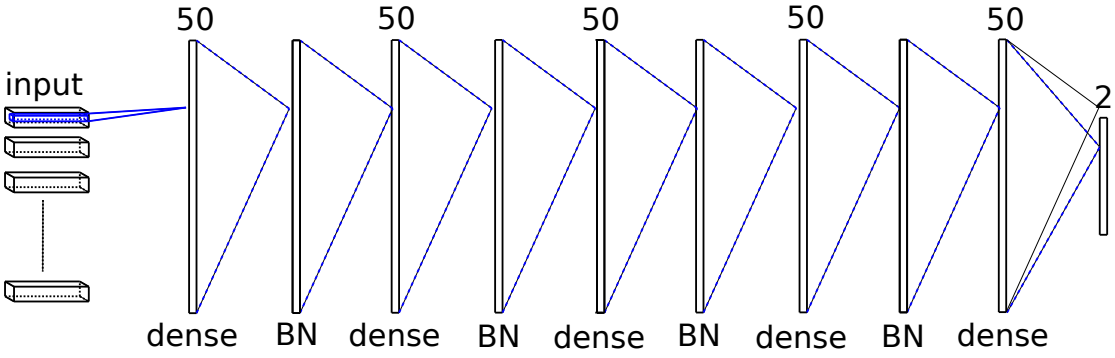


Fig. 7.3 Overview of the proposed MLP-based fusion scheme: 1 input layer followed by 4 pairs of dense/batch normalization (BN) layers, 1 dense layer, and 1 single-layer linear perceptron used for predicting the final interestingness score [6].

with various architectures and parameters, we arrived at following configuration: a deep network composed of 10 layers with 5 dense layers activated by the ReLu function, each followed by a batch normalization layer (bn). The network uses the interestingness prediction scores of the inducers as input during the training phase to learn intricate joint decisions. Figure 7.3 depicts the network's topology.

7.1.6 Evaluation

We utilize two split scenarios: i) 75% training and 25% testing (*RSKF75*), and (ii) 50% training and 50% testing (*RSKF50*). Split samples are randomized, resulting in 100 partitions. The official metrics are calculated as averages of these splits.

7.1.7 Results and discussions

The results obtained with our proposed MLP-based system are analyzed and compared to those obtained with the state-of-the-art systems from the MediaEval benchmark and the literature. Overall, it is evident that the ensemble system outperform the individual inducers.

7.1.8 Conclusions

One major finding is that, regardless of how strong a system is, ensembling many systems, even with individual average performance, improves performance. We were able to increase performance nearly in every case after testing with numerous fusion approaches, including typical late fusion of system scores, boosting techniques that employ weak learners, and a proposed, deep MLP-based system fusion. The proposed MLP system improves image prediction by 105 percent over state-of-the-art results, increasing mAP from 0.156 to 0.3202, and video prediction by 184 percent, increasing mAP from 0.093 to 0.2646.

7.2 Violence detection

The VSD96 data set [5] is a freely accessible dataset and a standardized evaluation framework for detecting violent sequences in Cinematic and YouTube videos. This dataset was validated at the MediaEval Violent Scenes Detection tasks from 2011 to 2015.

My main contributions to this dataset are as follows: (i) analyzing the employed classification techniques and their capabilities on the *VSD* data set that allow for general trends to be deduced with respect to the best performing systems, and (ii) an in-depth analysis of the representative methods from the literature, trained and tested on the VSD96 data that has not been submitted to the MediaEval Affective Impact of Movies, Violent Scenes Detection campaign.

7.2.1 Dataset description

The VSD96 dataset consists of prominent Hollywood films that span a broad variety of genres, topics, and levels of violence, ranging from extremely violent to films with no violence at all.

7.2.2 Methods analysis

We analyze the performance of more than 240 systems deployed for VSD.The following clusters of methods formed: SVM accounting for 65%, Statistical approaches for 13%, Neural Networks for 8%, Hybrid approaches for 8%, Discriminant Analysis for 2%, Deep Neural Networks for 2%, Clustering for 1% and Unsupervised learning with less than 1%. Looking at the overall performance, we noticed that Hybrid methods achieved the best results - which could be because of the inherently multi-modal nature of videos. These methods are followed by NNs, and DNNs.

7.2.3 Benckmarking of the state-of-the-art methods

To have a complete analysis of the existing systems' performance, we provide an in-depth analysis of the representative methods from the literature, trained and tested on the VSD96 data. These were not submitted to the MediaEval benchmark, and most importantly, they were developed without any time constraints. The complete results are presented in the full manuscript.

7.2.4 Conclusions

We have provided an in-depth analysis of the crucial components of the VSD algorithms, by reviewing the capabilities and the evolution of the existing systems with the objective to offer a complete practitioner's guide for this task. We reviewed 236 systems that were submitted to MediaEval and selected 17 state-of-the-art systems from the literature that were tested on VSD96 data, which constitute a strong baseline. We analyzed the reliability of the annotations and system rankings, examined various aspects. e.g., overall trends and outliers, the prediction methods employed, and the possibility of aggregating the systems' outputs into an ad-hoc super system to achieve even greater performance.

7.3 Face identification

Face Verification with Challenging Imposters and Diversified Demographics (*FaVCI2D*)[1], [26] is a novel face identification dataset composed of freely distributable resources compliant with data protection regulations.

My main contributions to this dataset are as follows: (i) provide an analysis of state-of-the-art architectures on existing face identification datasets to validate the obtained results to the reported ones in the literature and the fact that the feature extractors to be used for validating the *FaVCI2D* are configured correctly and their further comparison is fair. and ii) provide state-of-the-art features extracted from the *FaVCI2D* data set.

[1] work published under revision

7.3.1 Dataset description

FaVCI2D includes identities from 153 countries.The total number of unique IDs is 52,411, with 12,468 of them being used in genuine pairs. The total number of images in the dataset is 64,879.

7.3.2 Results and discussions

We report the accuracy of feature extractors in various configurations of impostor pair setting in Table 7.5. The similarity between imposter pairs' IDs varies between 1 and random, which is the standard verification condition. The size of the pool of imposters is varies between 1,000 and 52,410, the total number of IDs in *FaVCI2D*. Globally, the highest performance is achieved with *insightface*, whereas the worst performance is achieved with *facenet*. The usage of challenging pairings considerably decreases performance.

Model	Similar = 1 Imposter pool size					Similar = 10 Imposter pool size					Similar = 100 Imposter pool size					Similar = random Imposter pool size				
	1000	5000	10000	30000	52410	1000	5000	10000	30000	52410	1000	5000	10000	30000	52410	1000	5000	10000	30000	52410
insightface	97.37 ± 0.03	96.76 ± 0.04	96.56 ± 0.03	96.07 ± 0.03	95.75 ± 0.0	98.04 ± 0.02	97.58 ± 0.05	97.34 ± 0.02	96.89 ± 0.05	96.64 ± 0.0	98.62 ± 0.02	98.22 ± 0.01	98.05 ± 0.03	97.77 ± 0.05	97.50 ± 0.0	98.81 ± 0.02	98.78 ± 0.01	98.82 ± 0.01	98.81 ± 0.03	98.82 ± 0.0
ir152	93.62 ± 0.03	92.10 ± 0.07	91.36 ± 0.05	90.08 ± 0.04	89.48 ± 0.0	95.52 ± 0.05	94.11 ± 0.08	93.43 ± 0.15	92.33 ± 0.05	91.84 ± 0.0	97.17 ± 0.02	96.13 ± 0.07	95.54 ± 0.02	94.64 ± 0.05	94.00 ± 0.0	97.65 ± 0.03	97.66 ± 0.03	97.65 ± 0.03	97.63 ± 0.04	97.64 ± 0.0
seqface	91.73 ± 0.15	89.46 ± 0.15	88.47 ± 0.16	86.65 ± 0.17	85.61 ± 0.0	94.58 ± 0.08	92.37 ± 0.11	91.31 ± 0.14	89.76 ± 0.16	88.99 ± 0.0	97.18 ± 0.06	95.55 ± 0.02	94.56 ± 0.10	93.04 ± 0.10	92.28 ± 0.0	98.04 ± 0.03	98.03 ± 0.04	98.00 ± 0.04	97.94 ± 0.02	98.06 ± 0.0
vgg	91.52 ± 0.15	89.01 ± 0.13	87.79 ± 0.08	86.00 ± 0.11	85.28 ± 0.0	94.44 ± 0.10	92.13 ± 0.07	90.89 ± 0.17	89.19 ± 0.10	88.29 ± 0.0	97.27 ± 0.09	95.33 ± 0.08	94.44 ± 0.07	92.85 ± 0.07	91.92 ± 0.0	98.37 ± 0.04	98.35 ± 0.06	98.31 ± 0.05	98.32 ± 0.06	98.42 ± 0.0
facenet	89.74 ± 0.14	86.90 ± 0.12	85.44 ± 0.13	83.48 ± 0.08	82.61 ± 0.00	93.51 ± 0.09	90.39 ± 0.04	89.13 ± 0.11	87.07 ± 0.12	86.06 ± 0.0	97.13 ± 0.05	94.79 ± 0.08	93.52 ± 0.12	91.28 ± 0.07	90.11 ± 0.00	98.37 ± 0.07	98.37 ± 0.06	98.36 ± 0.06	98.36 ± 0.06	98.39 ± 0.0

Table 7.5 Verification accuracy with various models and setups. "Similar" indicates the imposter identity's position in the ranked list of similar identities in relation to the reference identity in each impostor pair. Lower values indicatea more difficult verification setup. The"imposter pool size" specifies the number of distinct identities from which an imposter could be selected. Higher values indicate a more challenging verification setup. Each configuration was run five times and the average accuracy and associated standard deviation are reported.

7.3.3 Benckmarking of the state-of-the-art methods

The following models were used in experiments: *insightface* [9], based on ResNet-150, trained on MS-Celeb1M dataset using ArcFace loss; *ir152* [45], based on ResNet-152, trained on MS-Celeb1M dataset using Focal loss; *seqface* [15], based on ResNet-27 trained on MS-Celeb1M using the L2-SphereFace loss and fine-tuned on Celeb-Seq dataset; *vgg* [1], based on SE-ResNet-50 trained on MS-Celeb1M dataset and fine-tuned on VGGFace2 dataset using Softmax loss; *facenet* [27], based on Inception ResNet, trained on VGGFace2 dataset using SoftMax loss.

In Table 7.6, In the full manuscript, we present the results obtained with the five models on two consacrated datasets. LFW [24] and YTF [41]. When available, the original model performance is reported in parenthesis. The results reproduced here are coherent with the original ones. This finding validates the fact that the feature extractors are configured correctly and their further comparison is fair.

Model	Training Data	LFW	YTF
insightface	MS-Celeb1M-ArcFace	99.87 (99.80+)	97.94
ir152	MS-Celeb1M	99.76 (99.80)	97.50
seqface	MS-Celeb1M + Celeb-Seq	99.80 (99.80)	98.00 (98.00)
vgg	MS-Celeb1M + VGGFace2	99.40	96.78
pfe	MS-Celeb-1M-ArcFace	99.82	97.36
dream	MS-Celeb-1M	97.71	95.01
facenet	VGGFace2	99.55	95.12

Table 7.6 Accuracy (%) of tested methods on LFW and YTF.

7.3.4 Conclusions

With the recent General Data Protection Regulation (GDPR) directive in European Union (EU) regarding data protection, establishing a new person identification dataset is extremely laborious. There have been many questions raised by the research communities whether they can build large datasets such as the public MegaFace [22] or the private DeepFace or FaceNet, owned by Facebook, and Google, respectively without any infringement of the aforementioned directive. While recent performance reports give the impression that the task is nearly complete, our studies reveal that existing evaluation datasets were constructed utilizing two oversimplifying design decisions. One identified cause is that the traditional identity selection used to construct imposter pairs is insufficiently difficult since, in fact, verification is required to discover challenging imposters.

Chapter 8

General conclusions and perspectives

8.1 Conclusions

This chapter summarizes my personal contributions during my doctoral research program to the field of computer vision, machine learning, medical, and financial, to name a few, using deep learning-based methods.

Part I presents the domain of the thesis and the motivation behind this work, followed by a literature review that addresses the methods that underpin deep learning implementations. Part II convers all of my personal contributions. Concretely, Chapter 3 presents my contribution to the medical domain. Sections 3.1– 3.3 describe a series of approaches for tuberculosis multidrug resistance detection, tuberculosis type classification, automatic image captioning and scene understanding in medical images, and lipreading. Chapter 4 presents my contributions to the surveillance domain. Sections 4.1– 4.2 describe approaches for joint person detection and re-identification and

person search, respectively. Chapter 5 presents my contributions to ensemble learning. Section 5.1 include a novel deep ensembling architecture designed to discover patterns and correlations between the decisions of individual classifiers. Chapter 6 presents my contributions to the fintech domain. Section 6.1 presents a method for financial time series generation and prediction of performing and underperforming stocks. Finally, Chapter 7 presents my contribution to the analysis, generation or publicly releasing of a collection of datasets. Section 7.1 presents the *Interestingness*10*k* [6] dataset for interestingness prediction in multimedia data. Section 7.2 presents the *VSD*96 [5] dataset for violent scenes detection, and Section 7.3 presents the *FaVCI2D* [26] dataset for face verification with challenging imposters and diversified demographics. Furthermore, they provide insights, observations and recommendations concerning the development of systems to predict the relevant representations for each of the datasets.

8.2 Contributions

- [C3, J1] present a framework for financial time series generation and prediction of performing and underperforming stocks. My main contribution comprises of developing the prediction algorithm, including a clustering approach for the stock universe clusterisation, a deep stacked autoencoder, and a deep neural network for predicting performers and underperformers.
- [J3] presents a comprehensive study of the Interestingness10k data set, with the goal of predicting image and video interestingness. To the best of my knowledge, this is the most comprehensive literature survey on the prediction of media interestingness at publishing time. My contributions comprise of analyzing the employed machine learning techniques and their capabilities on the data set that allows for general trends to be deduced with respect to the best-performing systems, providing insights regarding the capabilities of the newest deep neural networks by analyzing the performance of state-of-the-art architectures on the data set, visualizing and analyzing how algorithms interpret interestingness, providing a series of recommendations in respect to systems performances, developing an ensemble method based on the runs submitted to the MediaEval Predicting Media Interestingness competition, and taking part in the annotation process of the data set.
- [C1] presents a new face verification data set with challenging imposters and diversified demographics called *FaVCI2D*. My contributions comprises of providing an analysis of state-of-the-art architectures on existing face identification datasets to validate the obtained results to the reported ones in the literature and the fact that the feature extractors to be used for validating the *FaVCI2D* are configured correctly and their further comparison is fair, and providing state-of-the-art features extracted from the *FaVCI2D* data set.

- [C7, C5, B1] present a novel deep ensembling architecture designed to discover patterns and correlations between the decisions of individual classifiers. My main contributions comprise the development of two deep neural networks tailored for ensemble learning and defining a set of rules to build network architectures in a progressive manner. Validation is conducted on various classification and regression tasks, namely the Arousal and Valence detection subtasks of the MediaEval 2018 Emotional Impact of Movies task (two-class regression), MediaEval 2017 image and video subtasks from the Predicting Media Interestingness task (regression and cassification), the 2015 MediaEval Violent Scenes Detection task (binary classification), Fear detection subtask of the MediaEval 2018 Emotional Impact of Movies task (binary classification) and ImageCLEF 2019 Medical Concept Detection task (multi-label classification). At publishing time, these methods outperformed literature state-of-the-art with a great margin on all the data sets.
- [B3,B2, B5 B4, C2, C9] present a description of a series of public benchmarking competitions for which I was a member of the organizing team.
- [C4] propose a novel data set for lip reading for Romanian language. My contribution comprises of i) analyzing and implementing good practices for training deep neural networks and tailored data preprocessing for visual lip reading, and ii) developed visual lip reading models for heterogeneous hardware using Graph Lowering techniques. To the best of our knowledge, we were the first research group to release a lip reading dataset and models for Romanian language.
- [J2] presents a comprehensive study of the VSD96 data set with the goal of detecting violent video content. To the best of my knowledge, this is the most exhaustive literature study on the prediction of violent events from both subjective and objective perspectives at the time of publication, addressing and analyzing over 250 machine learning techniques. My main contributions to this data set comprise analyzing the employed classification techniques and their capabilities on the VSD data set that allow for general trends to be deduced with respect to the best-performing systems, and an in-depth analysis of the representative methods from the literature, trained and tested on theVSD96 data that has not been submitted to the MediaEval Affective Impact of Movies–Violent Scenes Detection benchmarking campaign with the goal of assessing the advances of the domain and provide the usefully lessons for the future.
- In [C6], I proposed a deep neural network-based unified architecture for person search, with attention mechanisms driving the detection phase. At the time of publication, attention techniques had not been implemented at the detection stage in an end-to-end fashion for person search, to the best of our knowledge.
- In [C11], I proposed a joint person detection and re-identification deep neural network in video streams. At the time of publication, the results outperformed the state-of-the-art approaches on the tested data set.

- In [C13], I proposed a flexible deep CNN structure that takes an arbitrary number of hypotheses as input for training multilabel image categorization. Results are validated in the ImageCLEF 2017Captiom task.
- In [C12], I proposed two completely automated tuberculosis categorization techniques and one automatic predictor for estimating the likelihood of TB patients developing multi-drug resistant TB, all base on deep neural networks. Furthermore, I have proposed a transformation approach for 3D CT images that allows applying transfer learning from other domains using deep neural networks. Results are validated in the ImageCLEF 2017 tuberculosis tasks.

8.3 Publications

Book Chapters:

B1: M.G. Constantin, **L.-D. Ştefan**, B. Ionescu : Exploring Deep Fusion Ensembling for Automatic Visual Interestingness Prediction. In book Human Perception of Visual Information - Psychological and Computational Perspectives, Springer International Publishing, Eds. B. Ionescu, W. Bainbridge, N. Murray. DOI: https://doi.org/10.1007/978-3-030-81465-6, September 10, 2021.

B2: B. Ionescu, H. Müller, R. Péteri, ... , **L,-D. Ştefan**, ... , J. Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications, in Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), Bucharest, Romania, Springer Lecture Notes in Computer Science LNCS, September 21-24, 2021.

B3: B. Ionescu, H. Müller, R. Péteri, ..., **L.-D. Ştefan**, ..., A. Popescu: The 2021 ImageCLEF Benchmark: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications. In: Springer Lecture Notes in Computer Science, vol 12657, DOI: https://doi.org/10.1007/978-3-030-72240-1_72, ECIR 2021 Proceedings, March 30, 2021.

B4: B. Ionescu, H. Müller, R. Péteri,..., **L.D. Ştefan**, M.G. Constantin, "Overview of the ImageCLEF 2020: Multimedia Retrieval in Medical, Lifelogging, Nature, and Internet Applications", in Springer Lecture Notes in Computer Science, 12260, CLEF 2020 Proceedings, September 22-25, Thessaloniki, Greece, 2020.

B5: B. Ionescu, H. Müller, R. Péteri, ..., M. Dogariu, **L.-D. Ştefan**, M.G. Constantin: ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications. In Springer Lecture Notes in Computer Science, vol 12036, pp. 533-541, DOI: https://doi.org/10.1007/978-3-030-58219-7_22, ECIR 2020 Proceedings, April 14-17, Lisbon, Portugal, 2020.

Journal papers:

J1: *Paper under revision:* M. Dogariu, **L.-D. Ştefan**, B.-A. Boteanu, C. Lamba, B. Kim, B. Ionescu, Realistic financial time-series generation. In Transactions on Multimedia Computing Communications and Applications (TOMM), 2021, (Q1 journal article, **Impact Factor: 3.1**).

J2: M.G. Constantin, **L.-D. Ştefan**, B. Ionescu, C.-H. Demarty, M. Sjöberg, M. Schedl, G. Gravier: Affect in Multimedia: Benchmarking Violent Scenes Detection. IEEE Transactions on Affective Computing, DOI: https://doi.org/10.1109/TAFFC.2020.2986969, April 2020. (Q1 journal article, **Impact Factor: 10.5**).

J3: M.G. Constantin, **L.-D. Ştefan**, B. Ionescu, N.Q.K. Duong, C.-H. Demarty, M. Sjöberg : Visual Interestingness Prediction: A Benchmark Framework and Literature Review. International Journal of Computer Vision, DOI: https://doi.org/10.1007/s11263-021-01443-1, 2021, WOS:000620409500001 (Q1 journal article, **Impact Factor: 7.41**).

Conference Papers:

C1: *Paper under revision:* A. Popescu, **L.-D. Ştefan**, J. Deshayes-Chossart, and B. Ionescu. Faceverification with challenging imposters and diversified demographics. In IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2022.

C2: R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, **L.-D. Ştefan**, M. G. Constantin and B. Ionescu. Overview of the 2021 ImageCLEFdrawnUI Task: Detection and Recognition of Hand Drawn and Digital Website UIs, Conference and Labs of the Evaluation Forum (CLEF), 2021.

C3: M. Dogariu, **L.-D. Ştefan**, B.-A. Boteanu, C. Lamba, B. Ionescu, Realistic financial time series data generation via generative adversarial learning. In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6-11 June, Toronto, Ontario, Canada, 2021, *ISI indexed conference*.

C4: A.-C. Jitaru, **L.-D. Ştefan**, B. Ionescu, Toward Language-independent Lip Reading: A Transfer Learning Approach. In International Symposium on Signals, Circuits and Systems (ISSCS), (pp. 1-4). IEEE, 2021, *ISI indexed conference*.

C5: M.G. Constantin, **L.-D. Ştefan**, and B. Ionescu. DeepFusion: Deep Ensembles for Domain Independent System Fusion. In International Conference on Multimedia Modeling, pp. 240-252. Springer, Cham, 2021, *ISI indexed conference*.

C6: **L.-D. Ştefan**, Ş. Abdulamit, M. Dogariu, M.-G. Constantin, and B. Ionescu. Deep learning-based person search with visual attention embedding. In 2020 13th International Conference on Communications (COMM), pp. 303-308. IEEE, 2020, WOS:000612723900053, *ISI indexed conference*.

C7: **L.-D. Ştefan**, M. G. Constantin, and B. Ionescu. System Fusion with Deep Ensembles. In Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 256-260. 2020, *ACM indexed conference*.

C8: M. Dogariu, **L.-D. Ştefan**, M. G. Constantin, and B. Ionescu. Human-Object Interaction: Application to Abandoned Luggage Detection in Video Surveillance Scenarios. In 2020 13th International Conference on Communications (COMM), pp. 157-160. IEEE, 2020, WOS:000612723900028, *ISI indexed conference*.

C9: D. Fichou, R. Berari, P. Brie, M. Dogariu, **L.-D. Ştefan**, M.G. Constantin, and B. Ionescu. Overview of ImageCLEFdrawnUI 2020: the detection and recognition of hand drawn website UIs task. In CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece (September 22-25 2020). 2020.

C10: C.-A. Mitrea, M.G. Constantin, **L.-D Ştefan**, M. Ghenescu, and B. Ionescu. Little-big deep neural networks for embedded video surveillance. In 2018 International Conference on Communications (COMM), pp. 493-496. IEEE, 2018, WOS:000449526000092, *ISI indexed conference*.

C11: **L.-D. Ştefan**. I. Mironică, C.-A. Mitrea. and B.Ionescu. End to end very deep person re-identification. In 2017 International Symposium on Signals, Circuits and Systems (ISSCS) (pp. 1-4). IEEE, 2017, WOS:000425211500061, *ISI indexed conference*.

C12: **L.-D. Ştefan**, Y. D. Cid, O. A. Jiménez del Toro, B. Ionescu, and H. Müller. Finding and Classifying Tuberculosis Types for a Targeted Treatment: MedGIFT-UPB Participation in the ImageCLEF 2017 Tuberculosis Task. In CLEF (Working Notes). 2017.

C13: **L.-D. Ştefan**, B. Ionescu, and Henning Müller. Generating captions for medical images with a deep learning multi-hypothesis approach: MedGIFT–UPB Participation in the ImageCLEF 2017 Caption Task.

C14: A. Toma, **L.-D. Ştefan**, and B. Ionescu. UPB HES SO@ PlantCLEF 2017: Automatic Plant Image Identification using Transfer Learning via Convolutional Neural Networks. In CLEF (Working Notes). 2017.

Research Projects:

R1: 2020–2024: **researcher**, project H2020 AI4Media. "A European Excellence Centre for Media, Society and Democracy", owner CERTH, Greece, partner Polytechnic University of Bucharest, axis H2020 ICT-48-2020 / Towards a vibrant European network of AI excellence centres (budget 12M Eur).

R2: 2020–2022: **researcher**, project SMARTRetail. “Enhancing and Improving Customer Experience and Services in Supermarkets via SMART Artificial Intelligence Powered Systems”, owner Softrust Vision Analytics, partner Polytechnic University of Bucharest, funded by UEFISCDI, Industry Transfer Axis, grant PN-III-P2-2.1-PTE-2019-0055 (budget 340k Eur).

R3: 2020–2022: **researcher**, project GRAVI. “Virtual Guardian: Artificial Intelligence Powered Multi-Sensor System for Automatic Securing of Areas of Interest”, owner

Softrust Vision Analytics, partner Polytechnic University of Bucharest, funded by UEFISCDI, Industry Transfer Axis, grant PN-III-P2-2.1-PTE-2019-0570 (budget 350k Eur).

R4: 2020 March–2020 October: **researcher**, project Keysight 1. "Machine Learning Techniques for Generating Network Traffic Data", owner Polytechnic University of Bucharest, CAMPUS Research Institute, beneficiary Keysight Technologies Romania (budget 59k Eur).

R5: 2020 April–2020 July: **researcher**, project Hana 2. "Financial Data Augmentation and Forecasting Using Advanced AI Techniques", owner "Polytechnic" Research, Development and Innovation Institute, beneficiary Hana Institute of Technology, Republic of Korea (budget private).

R6: 2019 May–2019 December: **researcher**, project NXP 1. "RISC V-based Hardware-Software System for Machine Learning Applications", owner Polytechnic University of Bucharest, CAMPUS Research Institute, beneficiary NXP Semiconductors Romania (budget 136k Eur).

R7: 2017–2020: **researcher**, project SPIA-VA. "Technologies and Innovative Video Systems for Person Re-Identification and Analysis of Dissimulated Behavior", owner Polytechnic University of Bucharest, partners UTI Grup, Romanian Ministry of National Defence — Military Equipment and Technologies Research Agency, public beneficiary Protection and Guard Service, funded by UEFISCDI, Solutions Axis, grant 2SOL/2017 (budget 2.2M Eur).

R8: 2017–2018: **researcher**, project SPOTTER. "Real-time IP Camera-based Intelligent Video Surveillance Security System with DROP Retrieval", owner Polytechnic University of Bucharest, partner UTI Grup, funded by UEFISCDI, PED Axis, grant 30PED/2017 (budget 140k Eur).

R9: 2016–2018: **researcher**, project Erasmus+ CBHE, UMETECH. "University & Media Technology for Cultural Heritage", owner University of Florence, Italy, partner Polytechnic University of Bucharest (budget 900k Eur).

8.4 Future perspectives

At the moment, the majority of a deep learning practitioners' work involves manipulating data with Python scripts and then meticulously tweaking the topology and hyper-parameters of a deep network to obtain a functional model. An adventurous researcher may recourse to an off-the-shelve state-of-the-art model. Unfortunately, this is not the ideal configuration. Supervised deep learning can assist in this area. Solutions that handle most of the model parametric tweaking may become a powerful tool, as we demonstrated in our ensemble learning experiments. A more ambitious perspective seeks to discover a suitable architecture from scratch through reinforcement learning or genetic algorithms. Another practical approach for automatization is to learn model architecture

in conjunction with model weights. Because training a new model from scratch each time we experiment with a slightly different architecture is highly inefficient. An effective automatic system would evolve architectures concurrently with tuning model features via back-propagation on training data, eliminating all computational redundancy.

References

[1] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE.

[2] Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *ACCV*.

[3] Constantin, M. G., Ştefan, L.-D., and Ionescu, B. (2021a). Deepfusion: Deep ensembles for domain independent system fusion. In *International Conference on Multimedia Modeling*, pages 240–252. Springer.

[4] Constantin, M. G., Stefan, L.-D., and Ionescu, B. (2021b). Exploring Deep Fusion Ensembling for Automatic Visual Interestingness Prediction. In *Human Perception of Visual Information: Psychological and Computational Perspectives*. Springer.

[5] Constantin, M. G., Stefan, L. D., Ionescu, B., Demarty, C.-H., Sjoberg, M., Schedl, M., and Gravier, G. (2020). Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*.

[6] Constantin, M. G., Ştefan, L.-D., Ionescu, B., Duong, N. Q., Demarty, C.-H., and Sjöberg, M. (2021c). Visual interestingness prediction: A benchmark framework and literature review. *International Journal of Computer Vision*, pages 1–25.

[7] Dellandréa, E., Huigsloot, M., Chen, L., Baveye, Y., Xiao, Z., and Sjöberg, M. (2018). The mediaeval 2018 emotional impact of movies task. *In Proc. of MediaEval 2018 Workshop*.

[8] Demarty, C.-H., Sjöberg, M., Ionescu, B., Do, T.-T., Gygli, M., and Duong, N. (2017). Mediaeval 2017 predicting media interestingness task. In *MediaEval workshop*.

[9] Deng, J., Guo, J., Niannan, X., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.

[10] Dogariu, M., Ştefan, L.-D., Boteanu, B.-A., Lamba, C., and Ionescu, B. (2021). Towards realistic financial time series generation via generative adversarial learning. In *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE.

[11] Dogariu, M., Ştefan, L.-D., Boteanu, B.-A., Lamba, C., Kim, B., and Ionescu, B. (2022). Realistic financial time-series generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*. [under review].

[12] Ghadiyaram, D., Tran, D., and Mahajan, D. (2019). Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055.

[13] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[14] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[15] Hu, W., Huang, Y., Zhang, F., Li, R., Li, W., and Yuan, G. (2018). Seq-face: make full use of sequence information for face recognition. *arXiv preprint arXiv:1803.06524*.

[16] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[17] Ionescu, B., Ghenescu, M., Răstoceanu, F., Roman, R., and Buric, M. (2020). Artificial intelligence fights crime and terrorism at a new level. *IEEE MultiMedia*, 27(2):55–61.

[18] Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.-T., Cid, Y. D., Eickhoff, C., de Herrera, A. G. S., Gurrin, C., et al. (2017). Overview of imageclef 2017: Information extraction from images. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 315–337. Springer.

[19] Jitaru, A. C., Abdulamit, Ş., and Ionescu, B. (2020). Lrro: a lip reading data set for the under-resourced romanian language. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 267–272.

[20] Jitaru, A.-C., Ştefan, L.-D., and Ionescu, B. (2021). Toward language-independent lip reading: A transfer learning approach. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE.

[21] Kakushadze, Z. and Yu, W. (2016). Statistical industry classification. *Journal of Risk & Control*, 3(1):17–65.

[22] Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., and Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882.

[23] Kinlay, J. (2011). Can machine learning techniques be used to predict market direction? The 1,000,000 model test.

[24] Learned-Miller, G. B. H. E. (2014). Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst.

[25] Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159.

[26] Popescu, A., Ştefan, L.-D., Deshayes-Chossart, J., and Ionescu, B. (2022). Face verification with challenging imposters and diversified demographics. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE. [under review].

[27] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[28] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

[29] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[30] Sjöberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., Schedl, M., Demarty, C.-H., and Chen, L. (2015). The mediaeval 2015 affective impact of movies task. In *MediaEval*.

[31] Stafylakis, T. and Tzimiropoulos, G. (2017). Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*.

[32] Ştefan, L.-D., Abdulamit, Ş., Dogariu, M., Constantin, M. G., and Ionescu, B. (2020a). Deep learning-based person search with visual attention embedding. In *2020 13th International Conference on Communications (COMM)*, pages 303–308. IEEE.

[33] Stefan, L.-D., Cid, Y. D., del Toro, O. A. J., Ionescu, B., and Müller, H. (2017a). Finding and classifying tuberculosis types for a targeted treatment: Medgift-upb participation in the imageclef 2017 tuberculosis task. In *CLEF (Working Notes)*.

[34] Ştefan, L.-D., Constantin, M. G., and Ionescu, B. (2020b). System fusion with deep ensembles. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 256–260.

[35] Stefan, L.-D., Ionescu, B., and Müller, H. (2017b). Generating captions for medical images with a deep learning multi-hypothesis approach: Medgift–upb participation in the imageclef 2017 caption task.

[36] Ştefan, L.-D., Mironică, I., Mitrea, C. A., and Ionescu, B. (2017). End to end very deep person re-identification. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE.

[37] Sudhakaran, S., Escalera, S., and Lanz, O. (2020). Gate-shift networks for video action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[38] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[39] Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. (2019). Fixing the train-test resolution discrepancy. In *Advances in Neural Information Processing Systems*, pages 8250–8260.

[40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[41] Wolf, L., Hassner, T., and Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE.

[42] Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2017). Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424.

[43] Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., Long, K., Shan, S., and Chen, X. (2018). Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8.

[44] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

[45] Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J., et al. (2018). Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216.

[46] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

[47] Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., and Tian, Q. (2017). Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376.