This project has received funding from the European Union's Horizon 2020 research and innovation programme under **grant agreement No 780355**

FANDANGO DELIVERABLE

Deliverable No.:	D4.5
Deliverable Title:	Machine learnable scoring for fake news decision making prototypes
Project Acronym:	Fandango
Project Full Title:	FAke News discovery and propagation from big Data and artificial inteliGence Operations
Grant Agreement No.:	780355
Work Package No.:	4
Work Package Name:	Fake news identifiers, machine learning and data analytics
Responsible Author(s):	CERTH
Date:	30.10.2019
Status:	v1.0 - Final Document
Deliverable type:	DEMONSTRATOR
Distribution:	PUBLIC

Revision History

Version	Date	Modified by	Comments
V0.1	07.10.2019	Panagiotis Stalidis	Table of Contents
V0.2	14.10.2019	Efi Kafali	Content on multimodal fusion
V0.3	15.10.2019	Panagiotis Stalidis	Content on explainability
V0.4	17.10.2019	Thodoris Semertzidis	Content on introduction, executive summary and conclusions
V0.5	18.10.2019	Efi Kafali	Formating
V0.6	18.10.2019	Panagiotis Stalidis	Full draft
V0.7	29.10.2019	Nicolas Vretos	Internal Review
V1.0	31.10.2019	Panagiotis Stalidis	Quality check

Table of Contents

List of Figures

List of Tables

ABBREVIATIONS

ABBREVIATION	DESCRIPTION
H2020	Horizon 2020
EC	European Commission
WP	Work Package
EU	European Union

EXECUTIVE SUMMARY

This report documents the work done thus far regarding the development of a machine learnable scoring function for the task of fake news detection. It describes the state-of-the-art methods used for the combination of information from multiple modalities into a single decision function. Special consideration has been given to methods that permit predictions even in the case of missing modalities. Additionally, it describes state-of-the-art methodologies used to interpret the results provided by a machine learning classifier, allowing for an informed decision about the trustworthiness of an article. Methodologies to explain separately the prediction from each modality, as well as their combination to get a final estimation using a multimodal fusion classifier, have been put to test and findings are discussed in this deliverable. It is important to note that FANDANGO is providing an explanation and interpretation mechanism for the machine learning algorithms which is vital for the end-users to trust the platform. Such an ability to provide interpretable predictions can ultimately lead to the elimination of any bias that might be introduced in the system.

1. INTRODUCTION

The ultimate goal of the FANDANGO project is to provide a platform that can assist a user to easily discern the validity of a news claim. In order to make such a decision, all the parts that make up the claim must much to the truth. Therefore, all aspects of a news item have to be judged for validity and then the combined information has to be judged for consistency.

Task 4.1 is dedicated to extracting information from various sources and in different forms. Then Tasks 4.2 through 4.4 each describe the methodologies used to understand if any of the modalities carries misinformation in its own accord. Task 4.5 provides two additional modules for the FANDANGO platform.

The first module incorporates the information provided from the different modalities such as the entities mentioned in the textual modality and the locations that are being discussed, the location where the images were taken and the object that are being shown from the visual modality and the origin of the authors and publishers from the metadata modality in order to detect the consistency of the content of a news item.

The second module takes into account the predictions that are made for each modality of an article and the confidence of each of the classifiers so that a single aggregated prediction can be made for the article. This is accomplished by a machine learnable score function that weights data from the data lake to justify the fakeness or not of a news post. Both modules are described in the following section 2.

One of the major concerns of the end users of the FANDANGO platform is the extent of trust they can show to classification algorithms and machine learning models. Thus, one of the requirements that they set is that FANDANGO should not make a final decision about the trustworthiness of an article, this decision should be left to the human interpreter, but each tool should provide results that are rigorous and explainable.

Current state-of-the-art in machine learning techniques has been cluttered with deep neural networks for a wide range of applications such as image classification and natural language processing. While these techniques have achieved extremely high predictive accuracy in *in vitro* conditions, such as competitions, deep learning models are not easily interpreted and their predictions are hard to explain. In recent years a lot of effort has been dedicated to define the concepts of “understanding”, “interpreting” and “explaining” when it comes to deep neural networks and to provide tools for discerning what a “black box” machine learning model has actually learned. In section 3 we provide a description of these methods as well as how the FANDANGO platform provides interpretable and explainable results to the end users.

2. MULTIMODAL FUSION

In this section the state-of-the-art approaches on multimodal data fusion are described. Our proposed method is thoroughly analyzed, whereas our experiments on three different public datasets are also presented. The goal of the multimodal fusion in the FANDANGO project is the formation of one aggregated score, which will be formed by fusing the trustworthiness scores of text, image and metadata content. The initial training of the fusion score classifier comes from the gross annotation of articles based on the source publisher, as provided by the end users. A better training schedule will be provided by a per article annotation of news by the end users, provided by the end of the project.

2.1. INTRODUCTION

Lately, the industry of news has gone under quite significant changes. Unlike past years when news websites presented an event by using only text and images, photo captions, tags, keywords and video content are lately also used. In addition, the rise of social media has inserted a new component to the entire news industry, by bringing up the existence of multiple views of the same event, due to the plethora of social media users that may refer to it. As a result, a vast number of posts, tags, images, or videos can be used to describe the exact same fact, from a different aspect. Subsequently, this has also evoked the effortless spreading of fake news.

The different modalities that describe the same event, can be used for drawing a preliminary conclusion regarding its trustworthiness, where modalities include the image, video, audio, text or, in general, any data type that can be present on a news website in the context of describing an event. However, each modality carries different types of information and when individually analyzed, the resulting trustworthiness score is based only on that specific modality. Hence, the relationships between the different modalities, describing the same event, are not exploited and significant information that could contribute to the accuracy of the trustworthiness score may be dropped.

Multimodal fusion methods have been occasionally used, in order to take advantage of those more complex relationships between modalities referring to the same event. However, the computational complexity of such multimodal data fusion frameworks was a limiting factor, until the recent progress in processing power. Nevertheless, the boundless availability of information describing the same event has been the key factor to the wide use of such frameworks.

According to [1], there are two types of multimodal classification frameworks that are most frequently used. Regarding the first one, namely late fusion, the different modalities are classified individually and their classification outputs are then fused, to return a final decision, while in the second one, namely early fusion, a classifier is used to classify the different modalities and fuse them into a single entity.

An article can contain different modalities, i.e. text, image and video. In earlier tasks, predictions about the trustworthiness of each modality were obtained which with relevance to this task will be fused to produce a single aggregated prediction for the entire article. Furthermore, the topics, entities, extracted locations and the object that were detected on the image will also be fused to produce a score indicating their relevance.

The objective of multimodal classification in the context of the FANDANGO project is the assignment of a predefined label (class) to the entire article, based on its distinguishing modalities. Since the final decision is not of great help to the end user in the case of binary classification (trustworthy/not trustworthy), the multimodal classification result is returned as a trustworthiness score. This score is based on information retrieved by all the modalities, but also the relationships between them.

2.2. State-Of- The- Art

Many solutions have been proposed for the creation of an effective system, that can fuse the information coming from different data sources, to better represent one single phenomenon. Multimodal fusion has been used in various tasks, such as action or expression recognition, indoor localization and tracking, image/video classification, speech recognition, person/object/context recognition, medical diagnosis , etc. In this section, the state-of-the-art methods of multimodal fusion are described for a wide range of applications.

Regarding action detection, the approach in [5] proposed the fusion of skeleton and RGB signals of the Kinect sensor. Instead of a holistic approach, the researchers have utilized two supervised and unsupervised fusion methods to fuse groups of features that come from different channels, showing promising results when compared to depth-only action recognition approaches.

Multimodal fusion has also been used in the context of expression recognition. In [6] a deep multimodal fusion CNN is proposed for the recognition of 2D and 3D facial expressions, where a 3D face scan is represented as six different facial attribute maps. The multiple 2D maps are fed to a deep CNN for feature and multimodal fusion learning, resulting in the retrieval of more solid facial maps and thus, a more accurate prediction of facial expressions.

In reference to automatic recognition of anger as a sign of aggressive behaviour, Patwardhan et. al. in [23] have used different expression modalities obtained by Kinect cameras (joint position, movement, body posture, head gesture, face and speech) in a supervised learning classification task. The results imply that the multimodal fusion in combination with rule based features can be used for security systems, for the effective early recognition of aggressive behaviour.

In [7], a new method for indoor localization and tracking is implemented, where magnetic and visual sensing features are fused for the creation of a filtering framework to track users, with the higher aim to maximize the localization accuracy.

Guillaumin et. al. [8] have approached the standard image classification task by introducing multimodal fusion to a semi - supervised learning process. In more details, the researchers have fused keywords associated with image labels with the images, to examine whether other kinds of information can assist the learning process of an image classification task. A Multiple Learning Kernel classifier has been initially trained to score unlabeled images, while its outputs were later used for learning SVM and LSR classifiers on both the labeled and unlabeled images. The results have achieved competitive performance, validating that using keywords associated with the image labels can boost a multimodal semi-supervised learning where only a small amount of labeled images are provided.

Geng et. al. have proposed efficient heuristic methods for improving multimodal frameworks in video concept detection [9]. A two-stage semantic model is approached, with the first stage being a multimodal fusion model, used in the context of an unsupervised method to adapt the differences of element distributions, regarding training and testing domains. The multimodal fusion method proposed by this research (domain adaptive linear combination - DALC) has introduced the novelty of exploiting huge amounts of data between the training and the testing domain and was based on the distribution of the projection angles between the model parameters and the domain elements. The second stage includes a multimodal concept model (mechanical node equilibrium - NE), with the objective to form the concept correlations between elements and, finally, adapt each concept's score which are represented as nodes. The essential concept of this stage is the search of an equilibrium state in this dynamic node system. The

combination of these two stages (DALC+NE models) has achieved state-of-the-art performance in either the supervised or unsupervised semantic models.

With reference to speech recognition, the research of [10] has explored the effects of using multimodal fusion in speech recognition, in cases where an image provides information about the speech recording. The lattice algorithm is used to rescore the most likely sentences of a word level RNN, while the image is also used to augment the language model with most likely words. The outcome of this research proves that indeed the fused information provided by the image can improve speech recognition, compared to speech recognition models where only the speech recording is used, and that the use of a larger multimodal dataset can provide even more accurate accurate results. Another remarkable multimodal fusion automatic speech recognition approach was also proposed in [11], where a turbo-decoding forward-backward algorithm is applied to an audio-visual speech recognition task and outperforms other state-of- the-art hidden Markov or iterative models.

In the context of person/object recognition, based on human perception about objects while interacting with them, the researchers of [13] have approached the method of sensorimotor, by using a deep learning based multimodal fusion of the information sources for automatic 3D object recognition. Regarding person re-identification via camera sensors, Pala et. al. [14] have examined the outcomes of fusing clothing appearance artifacts with additional descriptive information, such as anthropometric measures that can be estimated by unconstrained poses, retrieved by RDB-D sensors. Moreover, a dissimilarity discovery framework is used for the creation and multimodal fusion of pedestrian appearance descriptors. The results of this research indicate that the multimodal fusion approach has increased effectively the identification of clothing descriptors.

In [15] Destelle et. al. propose a multi - sensor fusion method for a low cost human skeleton tracking. The researchers use position information collected with Kinect cameras in fusion with more accurate pose estimation information that were obtained by wearable inertial sensors. The outcome verifies that the introduction of a second source modality can supplement the limitations of the Kinect cameras and thus, the framework can be used for more accurate skeleton tracking.

Furthermore, in [16] a multimodal sentiment analysis is proposed, where extracted features from audio, visual and textual modalities are fused. In order for the fused modalities to be meaningful and thus, have an impact on the learning process, both feature and decision level methods related to the different modalities are used. Besides the visual features that undoubtedly have a positive impact on the overall sentiment analysis, textual analysis has been enhanced with sentic - computed based features, which has proven to be a quite significant improvement over the standard textual sentiment analysis process.

Multimodal fusion has also been applied to medical diagnosis approaches, as in [17], where Xu et. al. have developed a model to predict Alzheimer's disease (AD) and mild cognitive impairment (MCI), by following a multi-modality sparse representation approach. The multimodal fused medical images that were used have shown improvement on the classification of AD or MCI diseases. In the same context, Bernal et. al. in [18] have proposed a supervised multimodal fusion method for human action and activity recognition, which applies on the automated monitoring of medical processes. The different modalities that are fused in the context of this method include motion data acquired by wearable sensors and video data acquired by body - mounted cameras. The combination of multiple sensor data, but also the adaptive sampling technique to the video processing part of the multimodal framework show improvement over standard action and activity classification methods.

It is quite common that the data used for multimodal fusion tasks are acquired by a variety of sensors, including bio-sensors (smart-meter sensors or blood-pressure devices, fingerprints), other passive sensors (Kinect and other cameras, smart-phones) or user created information, such as text or tags.

In [21] fingerprint and Iris data have been fused as a more robust approach in the context of human biometrics identification. Similarly, the research of [22] proposed a multimodal fusion framework for the classification of Alzheimer's disease, where the complementary information regarding data from different modalities were provided by a nonlinear graph fusion process.

In addition to the already described approaches that obtain their data by the most widely used passive sensor, Kinect, other types of passive sensors are also used in multimodal fusion frameworks. Cricri et. al. have followed a multi - user and multimodal approach for the automatic editing and classification of sport videos [24]. For the extraction of domain knowledge, multiple users were capturing audio/visual content about a sport event with their mobile phones, while also some auxiliary sensor data were captured by accelerometers and magnetometers. The different modalities were separately analyzed and the multiple different analyses were fused for acquiring the sport type. The sensor data have been used for the extraction of more specific and discriminative spatio-temporal features about the content. This multimodal and multi - user adaptive fusion approach has outperformed other state-of-the-art classification approaches. In [25] a multimodal late fusion method is proposed, where the learning process is based on imperfect sensor data, approaching a quite realistic scenario when it comes to robotics related tasks. This method utilizes a two - stream CNN that learns how to fuse information of RGB and depth image information automatically before the classification and has achieved impressive results regarding RGB - D object recognition tasks. Similar data are also used in [26], a multimodal video approach for the prediction of human - gaze direction. Finally in [27], features from pre trained DCNNs are used in fusion with other visual descriptors and audio modality extracted by openSMILE for personality prediction based on facial emotion and other information that can be retrieved from images.

In reference to user created data that are used in multimodal fusion systems, Chen et. al. in [28] have approached the multi-label image classification task by proposing a new method, which assigns the labels of each group to a query image. The proposed method is an exclusive Lasso model, that incorporates label exclusive content provided by the user, to a linear representation. Moreover, a study in [30] explores the effect of different user - created textual, visual and multimodal modalities, to social event classification, with relevance to social - event classification.

There is a plethora of methods applying multimodal fusion on different levels, based on the stage at which the data fusion takes place (early fusion, late fusion, feature level fusion, decision level fusion). An extensive analysis performed in [1], categorizes the methods based on the diversity and the type of the multimodal data to be fused, while the in [3] the most challenging issues arising from the use of multimodal data fusion methods are described, including data imperfections caused by sensors, data outliers, time-invariant and varying with time data, different data preprocessing, different data dimensions.

Regarding multimodal data analysis, in [12] a framework for multimodal content retrieval is presented that supports retrieval of rich media objects as unified sets of different modalities (image, audio, 3D, video and text). An automatic weighting method is used for combining the heterogeneous similarities of a single modality to one global modality. Moreover, a multimodal space is constructed, where the correlations of different modalities on a semantic level are captured. This framework is also capable of handling external queries, by incorporating them to the composed multimodal space. In [20], Kalimeri et. al. present a multimodal fusion framework using the fusion of electroencephalography (EEG) and electrodermal activity (EDA) signals, for assessing the emotional and cognitive experience of blind and visually impaired people, when navigating in unfamiliar indoor environments. In addition, the researchers of [19] have proposed a method which sustains multimodal data retrieved by electroencephalography signals (EGG). EGG video and optical flow data have been used for training deep CNN and RNN architectures in the context of an EGG video classification task, where the EGG signals are converted to grayscale videos and are used in fusion

with the optic flows of the EGG video, for a better representation of the temporal information of an EGG video. This multimodal method has solved the problem of inadequate EGG datasets.

According to the level of fusion, most of the research methods, follow the feature - level fusion approach, i.e. concatenating the features and applying a classification method, (e.g. ([20], [21]), or the decision - level fusion, i.e. using the classification results for each modality to improve the results by aggregation (e.g. ([23], [26]). There are also approaches which use both feature and decision level fusion (hybrid fusion). These methods utilize the joint effect that they have on the final decision, rather that the relationship between the multiple different modalities.

Despite the plethora of applications where multimodal fusion has already been used, most of the research approaches do not make use of the relationship between the different modalities, in order to classify the multimodal input. An approach that takes advantage of the relationships across multiple modalities is in [31], where a temporal generative model learns shared representations across different modalities with time varying data, in the context of event detection and classification. However, [31] cannot handle missing modalities and thus, there is the essential need of generating them, by relying on information provided by the available modalities. Similarly,in [32] Sohn et. al. explore the topic of identifying which are some good associations between different modalities that can be used for performance improvement in multimodal data fusion systems. In this approach, information theory based measures are used to generate missing modalities. Therein, the concatenated modalities are used to train a single representation using neural networks. Finally, in [24] the different modality relationships are described by a feature weight vector, which is computed by the concatenation of the different modalities and unimodal classification.

The proposed method is based on and [29], which also deal with multimodal classification tasks. In [33], Li et. al., based on the idea that photos of the same object can be modeled as different modalities, approach the multimodal image classification task with a Multi-manifold sparse graph embedding algorithm (MSGE), which is capable of capturing multimodal multi-manifold structure. However, this approach only uses one data type, image, while the fact that multiple photos of the same object are considered as different modalities is restricting. Moreover, [33] cannot be applied on data where there are modalities of different size or type. Another graph-based classification approach has been studied in [34], where image understanding and feature learning are integrated into a joint learning framework, by the proposed Robust Structured Subspace Learning (RSSL) algorithm. This method gives decent performance regarding image tagging, but has not been used for multimodal classification tasks. In the same context, in [35] a new weakly-supervised metric learning was proposed, which is modeled as a deep learning framework. The learning process includes the joint exploitation of visual content and user tags or social images. However, this method has only been used for image retrieval and cannot be applied to a multimodal classification task. Our proposed method utilizes matrix factorization in multiple steps of the process, which has also been studied for image understanding in [36].

2.3. Methodology

The proposed framework that will be used for fusing the different modalities of an article consists of some discrete parts that are described in detail in this section. In the first place, a data normalization process is used for subtracting the mean value of the features of each modality. Next, Singular Value Decomposition (SVD) transformation is applied in order to convey all the modalities to the same space. The space where the modalities are conveyed is the space of one of the existing modalities.

The next step includes the mapping of the unimodal data-input into multiple representations in a new space. In the new space, the computed representation is based on the Truncated SVD transformation matrix [37] and the distance between data is calculated. The Truncated SVD is used in the means of

dimension reduction of the data. Moreover, with this transformation it is ensured that the loss of information regarding the data will be negligible.

The following step is important, since its objective is to find a proper metric to measure the distance between the different modalities. Hence, the transformation matrix is updated so that the distance between data of the same/different class (label) is decreased/increased respectively [29]. The different modalities can be mapped into a space where the distance between data of the same class is small, while in the opposite case the distance is large. This mapping is then used for measuring the distances between the modalities. Each update of the matrix triggers an update to the winning mapping, which better fits to the general objective.

Finally, the multiple transformed modalities are passed separately to the graph-based method, which is utilized in order to preserve the distances between the different modalities. By using the graph, the target representation is derived in the new space, where any classification method can be applied on the individual modalities.

The steps of the proposed method can be summarized in the following list, while for a more comprehensive analysis they are also illustrated in Figure 1.

1. Normalization of the extracted features for all the modalities
2. Mapping of the extracted features of all modalities to the feature space of one single modality.
3. Initialization of the first transformation matrix V.
4. Update of the first transformation matrix based on the computed distance between modalities.
5. Computation of the final transformation matrix using a graph-based method.

2.3.1. NOTATIONS

We refer with X to a multimodal dataset, consisting of m modalities of n labeled samples, with labels l_i, i ∈ {1, . . . , c}. l_i stands for the class of the i^{th} sample, while c is the number of classes. The representation of each sample of the modality j ∈ {1, 2, . . . , m} is represented by the vector $x_{i,j} \in R^{dj}$ where d_j is the dimension of the sample representation for the modality j. Each sample can be represented by one vector x_i by concatenating its single-modality representations. Namely:

$$\mathbf{x}_i = [\mathbf{x}_{i,1}| \dots |\mathbf{x}_{i,j}], j = 1, 2, \dots, m \quad (1)$$

where x_i is of size [1 × d], $d = \sum_{j=1}^{m} d_j$. The entire dataset is represented by the matrix X of size n × m. Each row of X is the sample representation x_i namely:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \left[\begin{array}{c|c|c} \mathbf{x}_{1,1} & \dots & \mathbf{x}_{1,m} \\ \vdots & & \vdots \\ \mathbf{x}_{n,1} & \dots & \mathbf{x}_{n,m} \end{array}\right] = [\mathbf{X}_1| \dots |\mathbf{X}_m] \quad (2)$$

In details, the k^{th} value of the j^{th} modality of sample i is located at the i^{th} row and the (f (j) + k)th column of X, where $f(j) = \sum_{q=1}^{j=1} d_j$. X_j denotes the n × d j sub-matrix of X that contains the j_{th} modality features of all n samples.

2.3.2. Data Preprocessing

Regarding data preprocessing, the mean value of each submatrix X_j of the multimodal dataset is calculated and subtracted per dimension. In details, the mean value can be calculated by:

$$\boldsymbol{\mu}_j = [\mu_{j,1}, \ldots, \mu_{j,d_j}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,j} \tag{3}$$

and $z_{i,j} = x_{i,j} - \mu_j$, with Z being X after the normalization process.

The following step includes the transformation of all the modalities j, $w \in \{1, 2, \ldots, m\}$, $j \neq w$ in such a way that the distance $||Z_w - R_j Z_j||$ can be minimized. For each modality j,$w \in \{1, 2, \ldots, m\}$, $j \neq w$ the rotation matrices R_j are calculated. Except for the w^{th} modality where $R_w = I_{dw}$, with I_{dw} being the $[d_w \times d_w]$ identity matrix, the rest of the modalities are transformed. The solution of this minimization problem is obtained using SVD over the covariance matrix $COV(Z_q , Z_w) = Z^T_w Z_j$.

$$\begin{aligned}\mathbf{Z}_w^T\mathbf{Z}_j &= \mathbf{U}_{\mathfrak{R}}\boldsymbol{\Sigma}\mathbf{V}_{\mathfrak{R}}^T \\ \mathfrak{R} &= \mathbf{V}_{\mathfrak{R}}\mathbf{I}\mathbf{U}_{\mathfrak{R}}^T\end{aligned} \tag{4}$$

where V_R is a matrix of size [d j × d j], I is of size [d j × d w] with its elements being equal to Kronecker delta ($I_{ij} = \delta_{ij}$), and U_R is of size [d w × d w].

From now on, X will indicate the transformed modalities with $X_i = R_i \cdot Z_i$, $\forall i \in \{1, 2, \ldots, m\}$ are then concatenated as in (2) resulting in matrix X.

2.3.3. Initial Transformation

Truncated SVD Decomposition is applied on the dataset for s largest singular values so that $X \approx U \Sigma V^T$. V is the initial transformation matrix of size [d × s] where s is manually selected. In the case that s < d, dimensionality reduction of the dataset is also achieved. V can be further updated by considering the specific class information.

An iterative update algorithm is applied on the transformation matrix V (Algorithm 1). Considering x_α and x_β are two random samples, for all pairs x_α and x_β, the representations h_α and h_β of the two respective samples and α , $\beta \in \{1, 2, \ldots, n\}$, $\alpha \neq \beta$ are computed:

$$\mathbf{h}_\alpha = \mathbf{x}_\alpha \cdot \mathbf{V} \text{ and } \mathbf{h}_\beta = \mathbf{x}_\beta \cdot \mathbf{V} \tag{5}$$

Next, the updated values for the transformation matrix V are computed:

$$\begin{aligned}\mathbf{u}_\alpha &= \mathbf{x}_\alpha^T \cdot (\mathbf{h}_\alpha - \mathbf{h}_\beta) \cdot 2 \cdot C_1 \cdot st \\ \text{and } \mathbf{u}_\beta &= \mathbf{x}_\beta^T \cdot (\mathbf{h}_\beta - \mathbf{h}_\alpha) \cdot 2 \cdot C_2 \cdot st\end{aligned} \tag{6}$$

where C_1 and C_2 are parameters used for balancing the significance between the update matrices u_α and u_β and *st* is the step-size parameter that determines how quickly the transformation matrix will converge to

the final matrix. For the pairs of α and β belonging to the same cluster ($l_\alpha = l_\beta$) the transformation matrix V is updated as:

$$\mathbf{V} := \mathbf{V} - (\mathbf{u}_\alpha + \mathbf{u}_\beta) \tag{7}$$

while for the pairs of α and β with $l_\alpha \neq l_\beta$, V is updated if the euclidean distance between h_α and h_β is below a threshold :

$$\mathbf{V} := \mathbf{V} + (\mathbf{u}_\alpha + \mathbf{u}_\beta) \text{ , iff } \|\mathbf{h}_\alpha - \mathbf{h}_\beta\|_2 \leq \epsilon \tag{8}$$

This process is repeated for r iterations, until V matrix represents the similarities between modalities in the best way possible. The amount of iterations is predefined. For the bimodal case where there are two modalities (m = 2), the described method is illustrated in Figure 1.

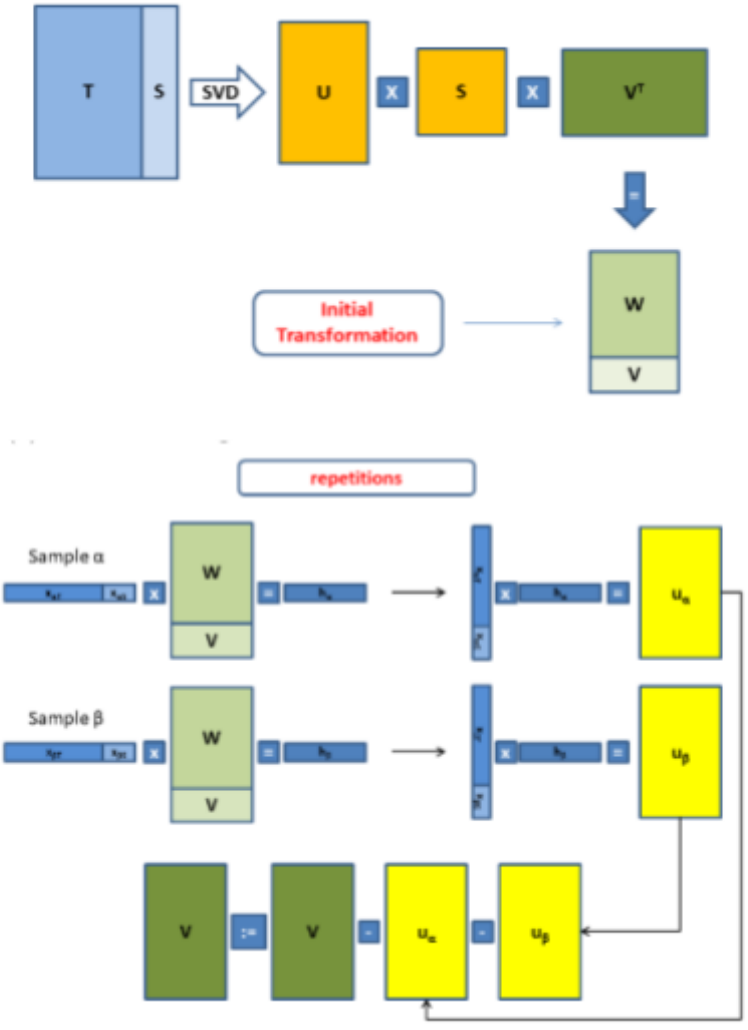

Figure 1 Transforming data for the initial transformation matrix V. T and S correspond to the first and second modalities accordingly. a) Initialization of the first transformation matrix V, using SVD b) The repetitive process of updating the transformation matrix

2.3.4. RESHAPING THE DATASET

The resulting transformation matrix V is obtained after r iterations. V contains m concatenated matrices V_j:

$$\mathbf{V} = \left[\mathbf{V}_1 \cdots \mathbf{V}_m\right]^T \tag{9}$$

Each V_j matrix represents the j^{th} modality of the dataset and is of size $[d_j \times s]$. s stands for the number of the largest singular values of the truncated SVD.

Next, the dataset is divided into m parts, with each part representing one individual modality. After reshaping, the dataset can be written as :

$$\mathbf{X}' = \left[\mathbf{X}'_1 \cdots \mathbf{X}'_m\right] \tag{10}$$

Algorithm 1 The SGD Updating Algorithm

Initialize $\mathbf{V} = \mathbf{V}_0$ by applying SVD
repeat
  repeat
    given a pair of samples (α, β) compute:
    $\mathbf{h}_\alpha$ and $\mathbf{h}_\beta$
    $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$
    if α and β similar $(l_\alpha = l_\beta)$ **then**
      $\mathbf{V} := \mathbf{V} - (\mathbf{u}_\alpha + \mathbf{u}_\beta)$
    else
      if $\|\mathbf{h}_\alpha - \mathbf{h}_\beta\|_2 \leq \epsilon$ **then**
        $\mathbf{V} := \mathbf{V} + (\mathbf{u}_\alpha + \mathbf{u}_\beta)$
      end if
    end if
  until a predefined number of pairs
until predefined number of iterations

where $X'_j = X_j \cdot V_j$ is a sub-matrix of X', containing the transformation of the j^{th} modality of all samples. X'_j is of size $[n \times s]$, hence the size of X' is $[n' \times s]$, where $n' = n \cdot m$. This is because a standard matrix multiplication gives:

$$\mathbf{X} \cdot \mathbf{V} = \mathbf{X}_1 \cdot \mathbf{V}_1 + \cdots + \mathbf{X}_m \cdot \mathbf{V}_m = \sum_{j=1}^{m} \mathbf{X}_j \cdot \mathbf{V}_j \tag{11}$$

This far, our method is based on the proposed method of [29], where a single representation for the concatenation of the two modalities is computed. We extend [29] and measure the distance between the different objects consisting of all modalities at this representation. We assume that we can keep the parts of the V that correspond to the j^{th} modality in order to map each modality to the new feature space individually to finally also measure the distance between different modalities in this new space.

2.3.5. Creating the Graphs

This phase includes the construction of two graphs, which are constructed for the reshaped dataset X'. The constructed graphs include a between-class graph $\{G_b , W_b\}$ and a within-class graph $\{G_w , W_w\}$, with W_b and W_w being the weight matrices of the two graphs, accordingly.

In order to create meaningful edges so that the connections between the different nodes of the graphs can be established, the kernel-based distances that are considered as similarity of the nodes are used. This is based on the hypothesis that the data lie on multiple manifolds in order to put constraints for the connections between the nodes. Furthermore, the evaluation of the kernel-based distances will be used as.

Given n' data samples $X' = [x'_1 , \ldots , x'_{n'}]^T$ belonging to c different classes, the data-points are separated into $\rho = c \cdot m$ modality manifolds $M = \{M_1 , \ldots M_\rho\}$, where $M_{j,l}$ is defined as the j^{th} modality-manifold-fragment (MMF) of the l^{th} class.

Regarding the construction of the within-class graph $\{G_w , W_w\}$, it is based on the MMF inner structure. Thus, $\forall x_i \in M_{l,k}$ $(l = 1, \ldots , \rho , k = 1, \ldots , c)$ x_i is connected with all $x_j \in Q \subset M^i_{l,k}$, where $|Q| = k$, $d(x_i , x_j) < d(x_i , x_k)$, $\forall j \in Q$, $k \in M^i_{l,k} - Q$, and $M^i_{l,k} \subset M_{l,k}$ with all the elements of $M_{l,k}$. The elements that have been connected with x_i in previous steps and x_i itself are excluded. By way of explanation, for all vertices that belong to the same MMF (the samples that belong to the same class and the same modality, $x'_{i'} \in M^{M_{j,l}}$), an edge is added between $x'_{i'p}$ and $x'_{i'q}$ is among the k nearest-neighbors of $x'_{i'p}$. In cases where an edge is already present from a previous iteration, we select the next nearest-neighbor.

The described process is performed for each MMF. In the following step, edges are added between different MMFs that correspond to the same class using the following procedure:

$$\forall \text{ MMF } M_{i,k} \text{, connect } M_{i,k} \text{ with } M_{j,k} \text{ if } x_a \in M_{j,k} : d(x_a , x_b) < d(x_c , x_d) \; \forall x_a , x_c \in M_{i,k} \text{ and } x_b , x_d \in M_{l,k} .$$

Thus, an edge is added between the two closest vertices belonging to different MMFs. Two MMFs are considered connected if any two of their vertices are connected.

For the between-class graph $\{G_b , W_b\}$, the MMF inner connections are also used as described earlier. Next, edges are added between different MMFs that correspond to the same modality. The same restrictions as set in the within-class graph construction are also applied to the way that graph vertices are connected.

2.3.6. Graph Output

The adjacency matrices A_w and A_b are obtained from the created graphs that were described in the previous subsection. W_w and W_b weight matrices are calculated by using the heat kernel according to which the weights of an edge between two vertices x_i and x_j are given by:

$$\mathbf{W}_{i,j} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{h}} \quad (12)$$

The two laplacian matrices L_w and L_b then computed as follows:

$$\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w \text{ and } \mathbf{L}_b = \mathbf{D}_b - \mathbf{W}_b \quad (13)$$

D_w , D_b are the diagonal matrices where the elements of the diagonals are the row-sums of the weight matrix, namely $D_{w\,i,i} = \sum_j W_{i,j}$.

2.3.7. Final Data Representation

As described in detail in [33], the objective is to minimize the following three quantities:

1. $\mathbf{Y}^T\mathbf{L}_w\mathbf{Y}$ s.t $\mathbf{Y}^T\mathbf{L}_b\mathbf{Y} = \mathbf{I}$
2. $\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2$
3. $\|\mathbf{A}\|_{2,1}$

The final objective is the minimization of the following function, i.e. the minimization of the weighted sum of the three quantities above:

$$\min(\mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1}) \text{ s.t. } \mathbf{Y}^T\mathbf{L}_b\mathbf{Y} = I \quad (14)$$

where ϖ and σ are two parameters responsible for balancing, A is a transformation matrix, and

$$\|A\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{p}([\mathbf{A}]_{ij})^2}$$

or zero otherwise.

If F is the objective function we get:

$$\mathcal{F} = \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1} \quad (15)$$

By differentiating F with respect to A, setting it to zero and solving for A, we get:

$$\mathbf{A} = (\mathbf{X}\mathbf{X}^T + \sigma\boldsymbol{\Delta}/2\varpi)^{-1}\mathbf{X}\mathbf{Y} = \hat{\mathbf{A}}\mathbf{Y} \quad (16)$$

where Δ is a diagonal matrix whose i^{th} diagonal element $\Delta_{i,i}$ equals to $(\|\boldsymbol{\alpha}_i\|)^{-1}$ only when $\alpha \neq 0$ (α i is the i th row vector of A). Here we result in a different equation than the corresponding one in [33].

$$\begin{aligned}\mathcal{F} &= \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1}) \\ &= \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi(\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A} - 2\mathbf{A}^T\mathbf{X}\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) + \sigma\mathbf{A}^T\mathbf{\Delta}\mathbf{A} \\ &= \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi(\mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{X}\mathbf{X}^T\hat{\mathbf{A}}\mathbf{Y} - 2\mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{X}\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) + \\ &\quad + \sigma\mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{\Delta}\hat{\mathbf{A}}\mathbf{Y} \\ &= \mathbf{Y}^T[\mathbf{L}_w + \varpi(\hat{\mathbf{A}}^T\mathbf{X}\mathbf{X}^T\hat{\mathbf{A}} - 2\hat{\mathbf{A}}^T\mathbf{X} + \mathbf{I}) + \sigma\hat{\mathbf{A}}^T\mathbf{\Delta}\hat{\mathbf{A}}]\mathbf{Y}\end{aligned} \tag{17}$$

If we set: $\mathcal{L} = \mathbf{L}_w + \varpi(\hat{\mathbf{A}}^T\mathbf{X}\mathbf{X}^T\hat{\mathbf{A}} - 2\hat{\mathbf{A}}^T\mathbf{X} + \mathbf{I}) + \sigma\hat{\mathbf{A}}^T\mathbf{\Delta}\hat{\mathbf{A}}$ the minimization function can be re-written as:

$$\min(\mathbf{Y}^T\mathcal{L}\mathbf{Y}) \text{ s.t. } \mathbf{Y}^T\mathbf{L}_b\mathbf{Y} = \mathbf{I} \tag{18}$$

According to the Lagrangian method, the optimization problem can be solved by computing the eigenvectors corresponding to the l smallest eigenvalues of the following generalized eigenvector problem:

$$\mathcal{L}\mathbf{Y} = \lambda\mathbf{L}_b\mathbf{Y} \tag{19}$$

The optimization process is described in detail in Algorithm 2:

Algorithm 2 Optimization Algorithm

Initialize $\mathbf{\Delta_0} = \mathbf{I}$, $t = 0$, ϖ, σ

repeat

compute $\mathbf{Y}_t$ by solving $\mathcal{L}\mathbf{Y} = \lambda\mathbf{L}_b\mathbf{Y}$

update $\mathbf{A}_t$ using $\mathbf{A} = \left(\mathbf{X}\mathbf{X}^T + \frac{\sigma\mathbf{\Delta}}{2\varpi}\right)^{-1}\mathbf{X}\mathbf{Y}$

evaluate $\mathbf{\Delta}$ from $\mathbf{A}_t$

$t = t + 1$

until convergence or preset iterations

Consequently, the final representation of the pre-processed input data is

$$\mathbf{R} = \mathbf{x} \cdot \mathbf{V} \cdot \mathbf{A} \tag{20}$$

which can finally be fed into any single modality classification method.

In the testing phase, let O be an object to be classified that consists of $z \leq m$ modalities. Let $Z = \{z_1, \ldots, z_z\} \subset \{1, \ldots, m\}$ where Z is the set of the indices of the existing modalities of the object O and m the number of modalities of the training set. Then, O is represented as

$$\mathbf{O} = \frac{\sum_{i=1}^{z}(\mathbf{x}_{O,i} \cdot \mathbf{V}_{z_z} \cdot \mathbf{A})}{z} \quad (21)$$

where, $x_{O,i}$ is the representation of the i^{th} modality of O, $V_{z\,z}$ is the z_z^{th} submatrix of V. In a similar manner as in the training procedure phase, the input data are normalized and rotated by R in the testing phase as well. Thus, the μ_j vectors from the training step are kept and reused during the testing.

As a highlight, it should be noted that in cases where there are training samples with missing modalities, they can be included in the training procedure, skipping through some steps. If a modality of a sample in training is missing, the specific sample does not participate in the preprocessing modality-transformation phase or the metric learning part (in initialization only).Thus in the following step, the existing modalities (except the w_{th}) will be transformed using equation (4). The only pre-processing the specific sample is through is the normalization.

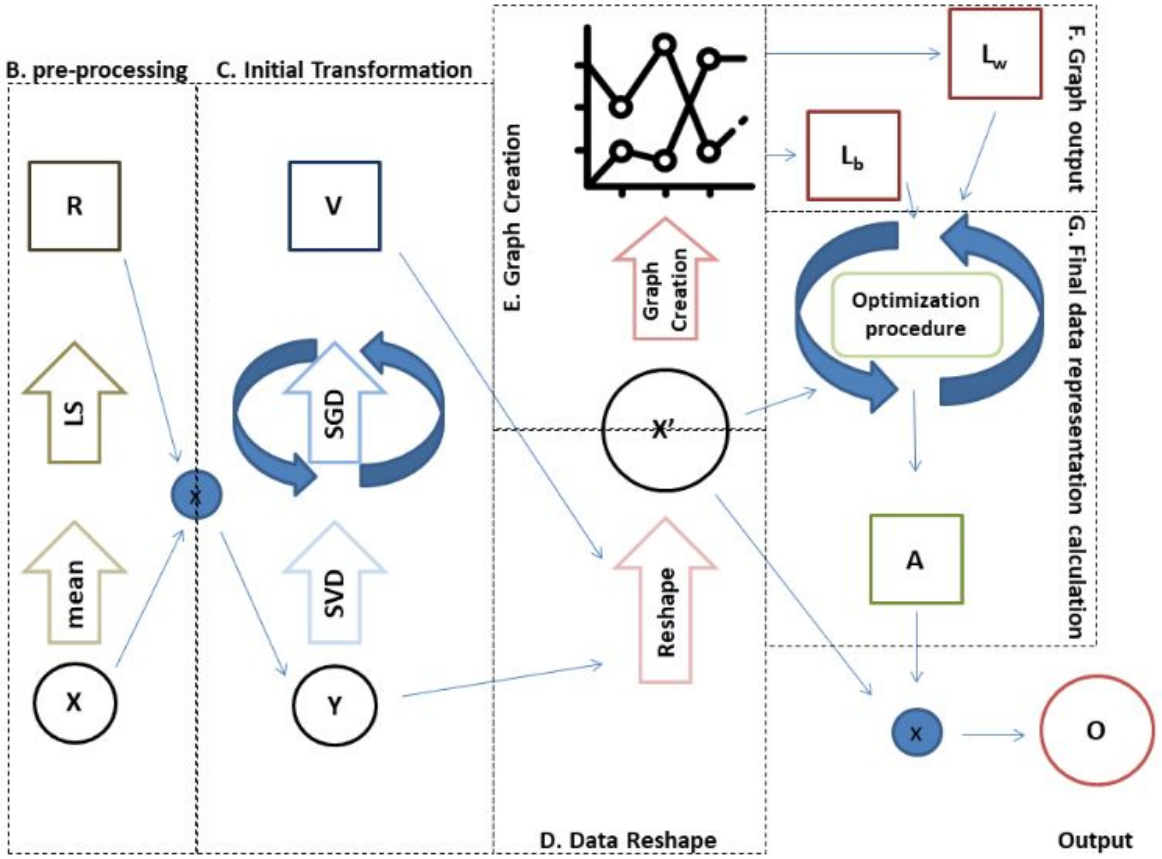

Figure 2 The different steps of the proposed method. The titles of the

parts are following the corresponding subsections in which the parts are described.

2.4. EXPERIMENTS

Experiments were conducted on three multimodal datasets: NUS-Wide [38], NTU RGB-D [39] and AV-Letters [40]. The following section includes descriptions and parameter settings of our experiments on the aforementioned datasets.

NUS-Wide samples consist of two different modalities, including images (six types of low-level features) and their associated tags from Flickr. We have used 1520 samples from the dataset (NUS-Wide 1.5K) and kept the bag-of-words feature based on SIFT descriptors.

NTU RGB-D dataset contains samples obtained by Microsoft Kinect cameras, with each sample representing a sequence of frames. One or two human bodies are recorded in each frame. For each human body, the 25 skeleton joints provided by the Kinect, are stored. For every joint, x, y, z (3D) coordinates of the joint, 2D (x, y) mapping of the corresponding depth frame, 2D (x, y) mapping of the corresponding RGB frame and 4D (w,x,y,z) orientation of the joint are provided. From the entire dataset, 500 samples are selected randomly.

Regarding the AV-Letters dataset, it consists of 780 samples, with each one of them containing a sequence of frames of various time lengths, including lip-movement while pronouncing a letter (video) and a sequence of Mel Frequency Cepstral Coefficients (MFCC) describing the audio of the respective letter.

For estimating the proper parameters for our approach, we have employed an exhaustive heuristic method. The search included the following parameters, which were described in previous sections:

- ϖ, σ, t,
- the number of singular values kept for the SVD step (sv),
- the total number of iterations of the updating algorithm (r),
- the number of eigenvalues kept in the Graph Optimization algorithm (l)
- and the trade-off parameters (C_1, C_2).

In Table 1, different values of the parameters used in experiments are shown. The optimal values are highlighted. Despite the fact that the three datasets we have experimented on are quite different, the hyperparameter selection shows that the optimal values were close for all three datasets.. This fact indicates that the proposed method achieves strong generalization across datasets.

2.4.1. NUS-WIDE 1.5K DATASET

Regarding NUS - WIDE 1.5K dataset, it also contains two different modalities. The first modality is a 1000d binary vector indicating the existence of 1000 tags on the image. The second modality is the probability of SIFT features that has been clustered into a bag-of-words of 1024 bins, to be found in the certain sample. 765 samples are used for training and 756 samples for testing.

The two modalities are concatenated following the steps described in section 2.3.1 resulting in a vector x_i of length d = 2024. Next, the SGD step is applied on V, for all the pairs of similar samples (totalling 9613 pairs), and for 10067 random dissimilar pairs.

In the following step, the dataset is reshaped. Each new samples has length sv = 60 and is labeled with the class and the modality it belongs to. The optimization process is repeated until convergence or until 200 iterations.

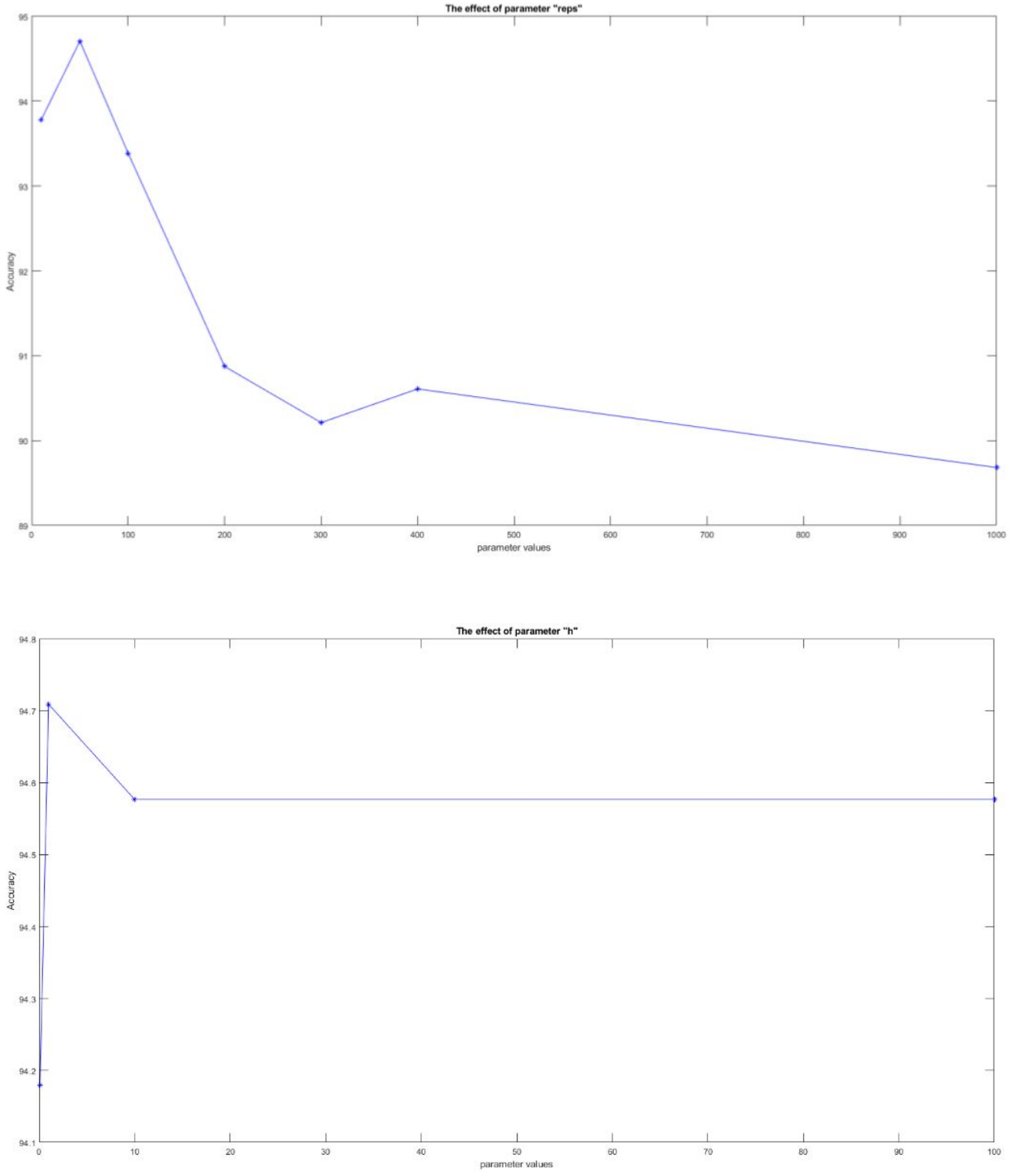

Figure 3 The effects of r and h parameters on multimodal accuracy. a) NUS-Wide 1.5k dataset: The effect of r parameter on multimodal accuracy b) NUS-Wide 1.5k dataset: The effect of h parameter on multimodal accuracy

We have applied 8-fold cross validation using multiple classifiers to compute the accuracy of each one of the different cases. The classifiers are trained with the mean of the resulting representations of each modality and with independent representations of each modality.

The threshold for updating V matrix in case of dissimilar samples, is set to 1. The results on NUS-Wide dataset are presented in Table 2.

	NUS	NTU	AV
ϖ	**1**, 10, 100, 1000	1, **10**, 100, 1000	**1**, 10, 100
σ	1, **10**, 100, 1000	1, **10**, 100, 1000	**1**, 10, 100
h	0.1, **1**, 10, 100	0.1, **1**, 10, 100	0.1, **1**, 10, 100
sv	10, 30, 50, **60**, 100	10, **30**, 50, 60, 100	10, 30, **50**, 60, 100
l	10, 30, 50, **60**, 100	10, **30**, 50, 60, 100	10, 30, **50**, 60, 100
r	50, **100**, 200, 400	50, **100**, 200, 400	**50**, 100, 200
C_1	**100**, 200, 300, 400	**100**, 200, 300, 400	100, 200, **300**, 400
C_2	200, 300, **400**, 500	**100**, 200, 300, 400	200, **300**, 400, 500

Table 1 The parameters used for the experiments

	Framework Parts				Classifiers Training	
	LS	**SVD**	**SGD**	**Graph**	**Modalities sep**	**Modalities av**
1		X		X	X	
2		X		X		X
3		X	X	X	X	
4		X	X	X		X
5	X	X		X	X	
6	X	X		X		X
7	X	X	X	X	**X**	
8	X	X	X	X		X

Table 2 Experimental implementations. The enumeration of cases are as described in section 2.4.4

	Multimodal CVA Accuracy		Per modality CVA Accuracy	
	av.	sep	Tags Mod	SIFT Mod
1	91.71%	48.98%	92.88%	5.13%
2	91.95%	48.82%	92.61%	4.28%
3	94.34%	50.53%	94.08%	6.97%
4	94.46%	49.70%	**94.60%**	4.87%
5	92.09%	49.48%	93.53%	5.42%
6	92.09%	49.60%	92.86%	5.81%
7	**94.74%**	**51.48%**	94.47%	**8.44%**
8	94.20%	49.64%	94.47%	4.87%

Table 3 The experimental implementations with tested accuracy and best classifiers and the experimental results per modality for NUS-Wide. The enumeration of cases are as described in section 4.4.

2.4.2. NTU RGB-D Dataset

Each frame sequence of the NTU RGB-D dataset is considered as a sample and the Bag of Words (BoW) method is applied on the data. We use a random 85% of the selected samples for training and the rest for testing. All the 25 × 3, 25 × 2, 25 × 2 and 25 × 4 features of all frames are concatenated into 4 matrices and then k-means is applied on them for 50 words (centers) each. Then, the probability of each word appearing in each sample is calculated. This procedure results in 4 sparse features of size (1 × 50), one for each sequence.

Since each sample is constructed as the concatenation of the 4 modalities feature vectors, its dimension is (1 × 200). The SGD step is applied on V for the same number of pairs of similar/dissimilar samples (1423/1423 pairs).

For the next step the dataset is reshaped at n 0 = n·m = 4 · 425 = 1700 samples. The new sample length after SVD is sv = 30. The results are presented in Table 4. The A and B rows in the table show the baseline. The baseline for comparison purposes is on the extracted features of the BoW, without any processing. The same classifiers and the same classification procedure is applied on the features to show the difference in terms of accuracy. Considering this dataset, results for one missing modality are presented in Table 5, and they are compared to the corresponding results of Table 4 under the av. column.

	MM CVA Accuracy		Per modality CVA Accuracy			
	av.	sep	Mod1	Mod2	Mod3	Mod4
$\mathcal{A}$	3.64%	57.45%	56.72%	59.15%	56.52%	57.35%
$\mathcal{B}$	61.54%	3.14%	3.23%	3.33%	5.00%	1.69%
1	50.00%	**62.50%**	**67.21%**	60.94%	62.50%	59.70%
2	60.00%	18.26%	11.86%	26.79%	32.79%	0.00%
3	55.00%	61.81%	60.61%	63.93%	**64.41%**	58.82%
4	58.21%	20.76%	18.75%	31.75%	24.14%	5.88%
5	3.77%	62.26%	60.94%	62.90%	60.61%	**64.62%**
6	57.58%	5.16%	2.08%	5.66%	5.88%	6.56%
7	40.98%	**62.50%**	60.00%	**65.00%**	64.06%	61.19%
8	**70.69%**	16.90%	18.97%	22.58%	19.61%	2.38%

Table 4: The experimental implementations for NTU.

	Overall output	Per missing modality CVA Accuracy			
		Mod1	Mod2	Mod3	Mod4
$\mathcal{B}$	61.54%	31.67%	31.58%	42.19%	28.07%
2	60.00%	40.91%	53.23%	50.72%	**50.77%**
4	58.21%	49.18%	50.00%	48.48%	47.69%
6	57.58%	9.43%	14.81%	17.46%	6.45%
8	**70.69%**	**51.61%**	**64.41%**	**60.32%**	43.94%

Table 5: The experimental implementations for NTU for 1 missing modality.

2.4.3. AV-Letters Dataset

A neural network is used for feature extraction. The network consists of two LSTM layers. The two modalities are passed through two similar neural networks separately. The outputs of the networks are feature vectors for each modality of each sample, one for audio and one for video. The length of each feature vector is [1 × 26] and represents the possibility of the sequence to be classified in one of the 26

classes. We used a random subset of 650 samples for training and the rest 130 for testing. The features of the two modalities are then concatenated resulting in a vector x_i of length d = 52. The SGD step is applied on V for 8160 pairs of similar samples, and for 8160 random number of pairs of dissimilar samples.

Then, the dataset is reshaped at n' = n·m = 2 · 650 = 1300 samples resulting (after SVD) in sample length equal to sv = 60. Finally, as in NUS-Wide dataset, a 5-fold cross validation has been applied. The results for AV-Letters dataset are presented in Table 6.

2.4.4. Results Interpretation

LS and SGD are procedures that are included/omitted in the experiments resulting in variations of the proposed framework as shown in the Framework Parts columns of Table 2.

As shown in (20), the training set elements that are fed into the classification method are the rows of R. More specifically, we calculate the representation for each modality R_j namely

$$\mathbf{R}_j = \mathbf{X}_j \cdot \mathbf{V}_j \cdot \mathbf{A} \tag{22}$$

where j = 1, 2, . . . , m. Each element $r_{i,j}$

$$\mathbf{R} = \left[\mathbf{R}_1 \middle| \dots \middle| \mathbf{R}_m \right] = \left[\begin{array}{c|c|c} \mathbf{r}_{1,1} & \dots & \mathbf{r}_{1,m} \\ \vdots & & \vdots \\ \mathbf{r}_{n,1} & \dots & \mathbf{r}_{n,m} \end{array} \right] \tag{23}$$

is the representation of the j th modality of the i th sample. Since all m modalities lie on the same space in the final representation, in the training procedure we can use from each sample either all the m l-d vectors (Separate training) or their average vector $R_i^{av} = \sum_{i=1}^{m} r_{i,j}$ (Average Training). Hence in the first case the number of training vectors per sample is m (in total m·n elements - Modalities sep column) while in the latter is one (in total n elements - Modalities av column), as indicated in Table 2 under Classifier Training columns.

	MM CVA Accuracy		Per Modality CVA Accuracy	
	av	sep	mod-A	mod-V
1	70.77%	46.54%	33.21%	**59.87%**
2	72.82%	41.60%	23.97%	59.23%
3	70.77%	46.67%	33.46%	**59.87%**
4	72.69%	41.41%	23.33%	59.49%
5	71.15%	**47.05%**	34.36%	59.74%
6	**73.21%**	41.15%	22.95%	59.36%
7	71.15%	**47.05%**	**34.49%**	59.62%
8	73.08%	41.03%	22.56%	59.49%

Table 6: The experimental implementations for AV-Letters.

Similarly, in the testing procedure the input consists of m l-d vectors for each sample. Thus, for the classification we can use the average of these m vectors (Multimodal CVA accuracy - sum column of Tables 2-6) or each modality representation separately (Per modality CVA accuracy columns), where CVA stands for cross validation average. In other words, on these columns, the method's accuracy is presented in the cases that all or only one modality is available. Column (Multimodal CVA accuracy - sep of Tables 2-6) equals the average of the m columns under Per modality CVA accuracy.

From Table 2 is shown that in NUS-Wide dataset, the Tag modality outperforms SIFT modality by far, with SIFT achieving very bad accuracy (< 10%). It can be also noted that the same happens for the AV-Letters dataset (Table 6), where Mod-V outperforms Mod-A.

During the data-preprocessing (Subsection 3.2), for the experiments on NUS-Wide and AV-Letters datasets, the modalities have been transformed to the first modality space (w = 1) as shown in Tables 2-6. This covers both possible cases, namely, in the NUS-Wide dataset we mapped the modality with the low accuracy to the one with the high accuracy, while in the AV-Letters we performed the inverse. We have also performed mappings to the other modalities (w 6 = 1) which resulted in similar with the initial case (w = 1) accuracies. On the contrary, the results on NTU-RGBD dataset are for modalities transformed to the 4 th modality as shown in Table 4. Even though the 4th modality shows the lowest accuracy on its own, when transforming all modalities to its space, the method shows significant improvement compared to transforming into the other modalities.

As expected, the results were better when the same training/testing object vector representation was used. Thus, the average of the modalities achieved better results in the classifier trained with the average of the modalities and vice versa.

The proposed method gives comparable results to the state-of-the-art methods. As can be seen in Tables 7 and 8, in AV-Letters database, our method surpasses the others by far for the multimodal case, even though for the case of the single-modal classification of AV-Letters is ranked last. In NUS-Wide database, our method surpasses previous state-of-the-art methods. In all cases the reader is referred to the referenced works for more results compared to other methods. The numerical results given in the Tables for the other methods, are taken by [41] and [29], respectively. For the NTU-RGBD dataset, we use as baseline for comparison the features extracted with BoW method. As it can be seen in Table 4, our method gives a significant boost to the accuracy, especially in the multimodal case that is shown in column under av. Moreover, in Table 5, the case where one modality is missing, and thus, the classification is done for the three remaining modalities. Only the Average training approach is presented.

However, the most significant contributions of our method is that it is universal, it can be applied to any kind of data, and most importantly, it can deal with the cases of missing modalities. For example, columns under Per modality CVA accuracy (Tables 2, 4, 6) illustrate the method's performance if only one modality is available during the method testing,and Table 5 shows the performance for one missing modality of NTU-RGBD dataset. Furthermore, objects with missing modalities can also be employed during training by following the Separate training approach.

	AV	A	V
MDAE [2]	62.90%	58.40%	62.10%
CRBM [31]	64.8%	61.2%	62.60%
RTMRBM [41]	66.04%	64.41%	64.63%
Proposed	**73.21%**	**34.49%**	**59.87%**

Table 7: Method performance in AV-Letters

Xie [29]	93.52%	Xing+Original	89.95%
ITML+Original	89.95%	Xing+MWH	89.95%
ITML+MWH	92.86%	MKE	80.56%
Proposed	**94.74%**		

Table 8: Method performance in NUS-Wide 1.5K. The un-cited methods' results are taken from [29]

2.5. IMPLEMENTATION

The data extracted from previous FANDANGO tasks can here be utilized for a more thorough analysis of a single media content. To start with, semantic information regarding the text of an article has been extracted in Task 4.1, including descriptors about mentioned entities, location and time. For the extraction of these data, each information is classified individually, resulting in a trustworthiness score. Currently, the trustworthiness score of the text modality of an article is based on the results of different NLP classification models, in the form of a weighted average score. In the context of Task 4.1, metadata descriptor vectors have also been extracted. These are based on information about the authors and publishers connected with the article and are used for determining their trustworthiness, as objects that exist in the FANDANGO data lake. A trustworthiness score of a publisher or an author is extracted by statistical models used to classify them as trustworthy/not trustworthy sources. Finally, in Task 4.3 descriptors about visual content (image/video) were extracted. Image segmentation models have been used for spotting potential manipulated areas of an image and based on the predicted mask, a trustworthiness score about the content is returned.

To determine the trustworthiness of a news article as a whole, different methods can be deployed. The naive approach would be the formation of one single trustworthiness score based on a weighted average of the computed scores for each modality. However, this approach would be way too simple for a critical case like the identification of fake news, since ambiguous relationships between the different descriptions could

be missed. Thus, having evaluated our method on the datasets described in section 2.4, we will apply it to the specific task of fake news identification. Since multiple modalities can be derived from an article/news post, their fused features can be crucial for shaping one single aggregated score, describing the overall media credibility. Moreover, a single trustworthiness score enriched with relational features between the different modalities will be a more meaningful and easier to interpret indicator for the end-user.

All the data will be preprocessed as described in the preprocessing step of our method. The text, metadata, image modalities will be transferred to a single space of one of those modalities. They will be fed to the graph method and their distance regarding their trustworthiness will be measured. The distance between different classes will be maximized, while the distance between similar classes will be minimized (trustworthy/not trustworthy). This will decide the relevance between different modalities belonging to the same article. As an example, the textual information of an article is describing an event by using suspicious words, meaning that its trustworthiness score based on the textual analysis is low. However, its visual content has a high trustworthiness score, since the image classifiers did not detect manipulated image areas. The distance between these two modalities will be maximized, since the NLP and segmentation classifiers resulted to different classes (not trustworthy/trustworthy).

At this point, the feature of our method regarding cases with missing modalities should be noted. Depending on the source, it is quite often that news articles do not always provide information about the multiple modalities. Articles with missing images, author or publisher information or metadata fields are often published in news sites. This factor plays an important role in the calculation of each modality's trustworthiness score and thus, the final aggregated score. Until recently, the solution to this problem would come by generating the missing modalities so that they are equipped with similar features as other existing modalities. Our method handles cases of missing modalities, since such modalities do not participate in the transformation phase we described earlier.

The goal of multimodal late fusion applied in the detection of fake news is the creation of one aggregated score that is both informative and easily deciphered by the end-user. Machine learning predictions can be confusing and easily misinterpreted, especially in a multileveled system like the FANDANGO platform. With the use of multimodal data classification, the end-user can consider one single score to begin with their analysis on a news article. This can save time, by the means of avoiding the interpretation of individual scores for each modality. Furthermore, a unified score describing a news article is a much simpler way to spot a suspicious news, than analyzing each component individually.

3. Explanation and Interpretation of Predictions

Machine learning is at the core of many recent advances in science and technology with SVM being a driving force in the previous decade and deep learning in this one. However, although these models reach impressive prediction accuracies in any task assigned to them, their nested non-linear structure makes them highly non-transparent. Because it is not clear what information from the input data makes them actually arrive at their decisions, these models are typically regarded as black boxes. But, whether humans are directly using machine learning classifiers as tools, or are deploying models within their products, in order to trust a model the decision process should be known and understood. Even more, if the users do not trust a model or a prediction, they will not use it.

Besides trust, there are more, perhaps even more important reasons to examine the reasons that a machine learning model reached a prediction such as learning from the system and complying with legislation. Because today's AI systems are trained with millions of examples, they may observe patterns in the data which are not accessible to humans, who are only capable of learning from a limited number of examples. If one has an explained decision process on can try to extract the distilled knowledge from the AI system and reach new insights on the process.

AI systems are affecting more and more areas of our daily life. Persons immediately affected by the decisions of an AI system may want to know why the systems has decided in this way. These concerns brought the European Union to adapt new regulations which implement a "right to explanation" whereby a user can ask for an explanation of an algorithmic decision that was made about her or him. Additionally, it is imperative to be able to assign responsibility when the systems makes a wrong decision.

These reasons force the development of two concepts: machine learning models must be interpretable and their predictions have to be explainable. In this context, an *interpretation* is a mapping of an abstract concept into a human understandable domain. Furthermore, an *explanation* is the collection of features of the interpretable domain that contributed for a given example to produce a decision. For example, an interpretable domain is an image where concepts like human, tree and dog can be depicted. The explanation would be a mask of which pixels contributed to the prediction of the class dog (Figure 4).

Figure 4 Example of an interpretable explanation for an image object detection model

Based on the definition of an interpretation, there is no constraint that the domain used to represent the explanation is the same as the domain of the input features to the system. This is especially important when the input domain is incomprehensible to a human; a new understandable domain can be selected. In natural language processing, for example, the selected classifier might be working with embeddings, which are representations of words and sentences into an n-dimensional hyperspace. But the interpretation could be the existence or absence of specific words in a bag-of-words setup.

In a bag-of-words setup, the domain size is limited by the dictionary of words, i.e. the set of unique words used in the text corpus. The size of the dictionary can in some cases reach several hundreds of thousands of words. An explanation that consists of the weighted appearance or absence of thousands of words, while understandable, is not very practical for a human. Therefore, this new requirement that further implies that the explanation should be easily understood introduces the concept of explanation *complexity*.

Another problem that arises, especially when the input and explanation domains are different, is that the explanation might not be completely faithful to the model. Fidelity is a measure of faithfulness of the explanation to the model. An explanation should be faithful to the model at least in the vicinity of the input under scrutiny, introducing the concept of *local fidelity*.

3.1. Existing Methods

Recent developments in machine learning like non-linear classifiers and deep learning models being so successful have led to algorithms that can not be properly understood even by practitioners of the discipline, let alone end users of the provided services. This has led to a number of methodologies to manage the understanding of machine learning algorithms and their predictions. Here we describe the 3 methodologies that are the most suitable to the FANDANGO project but the interested reader can follow recent research in Guidotti et.al. [42].

3.1.1. Layer-wise Relevance Propagation (LRP)

Layer-wise relevance propagation [43] is a general framework for decomposing predictions of modern AI systems such as feed forward neural networks, bag-of-words models, convolutional neural networks and recurrent neural networks. This method explains predictions relative to the state of maximum uncertainty, i.e. it identifies pixels which are pivotal for the model's prediction.

Mathematically, LRP redistributes the prediction $f(x)$ backwards using local redistribution rules, until it assigns a relevance score R_i to each input variable. The key property of the redistribution process is referred to as relevance conservation and can be summarized to:

$$\sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(x) \tag{24}$$

where i, j and k are input variables of different layers of the neural network. This property says that at every step of the redistribution process the total amount of relevance is conserved. The relevance scores R_j of each input variable j determines how much this variable has contributed to the prediction.

For a feed forward neural network, assuming x_j is the activation of neuron j at layer l, w_{jk} is the weight connecting neuron j to neuron k of layer $l+1$ and R_j and R_k are the relevance scores for neurons j of layer l and k of layer $l+1$ respectively, relevance R_j can be calculated by:

$$R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \epsilon} R_k \tag{25}$$

where ε is a small stabilization term to prevent division by zero. Based on equation (25), relevance is distributed proportionately from one layer to the next and relevance is higher where activation is higher, i.e. more activated neurons are more relevant, and where the strength of the connection is higher, i.e. more relevance flows through more prominent connection.

Equivalent relevance calculation can be achieved in other types of deep learning layers such as convolutional and recurrent by following the rules of information propagation ([43], [44], [45]).

This relevance calculation rule only takes into account positive relevance but there are many examples where we can consider an input having negative relevance. For example in an image depicting an animal with fangs this animal should not be classified as a sheep. An alternative relevance distribution rule, that adapts the process to this concept is:

$$R_j = \sum_k \left(\alpha \cdot \frac{(x_j w_{jk})^+}{\sum_j (x_j w_{jk})^+} - \beta \cdot \frac{(x_j w_{jk})^-}{\sum_j (x_j w_{jk})^-} \right) R_k \tag{26}$$

where $()^+$ and $()^-$ denote the positive and negative relevance parts. The conservation of relevance is enforced by the constraint $a - b = 1$. This methodology shows a close relation to Taylor decomposition, a general function analysis tool, as has been shown by [46] for the special case where $a = 1$.

Examples of explanations obtained with LRP, using different values of a and b, for predictions by a convolutional DNN of MNIST digits, are shown in Figure XX. When $a = 1$, the heatmaps contain only positive relevance which is spread along the contour of the digits in a fairly uniform manner.

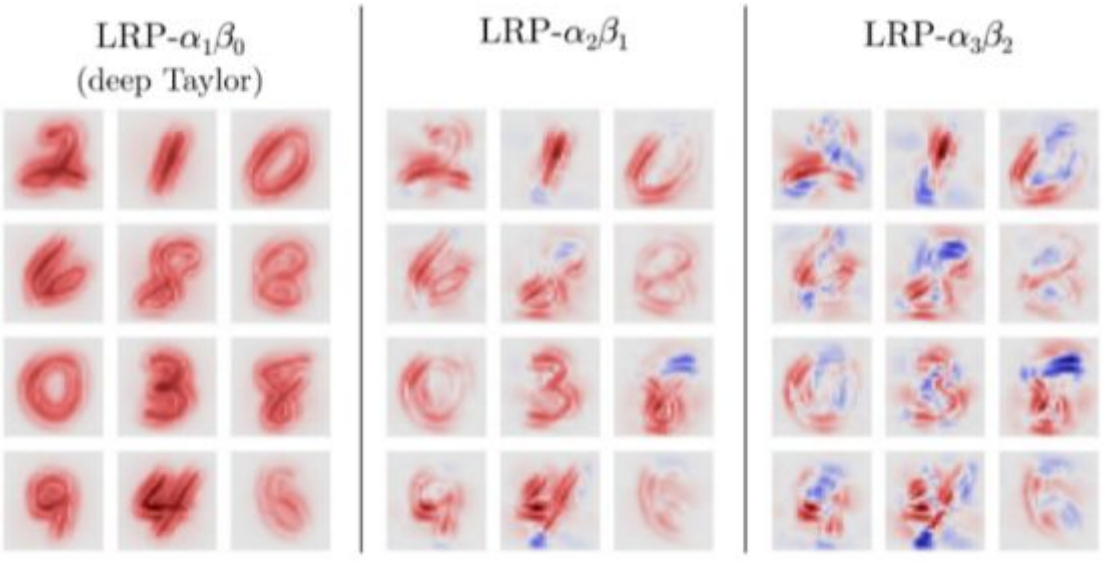

Figure 5 LRP explanations when choosing different LRP parameters a and b. Positive and negative relevance are shown in red and blue respectively [43].

When choosing $a = 2$, some regions in the images become negatively relevant, for example the upper region of the digit "8" in the last column of the third row. Setting an even higher value of $a = 3$, negative

relevance appears as if it is distributed in a seemingly random fashion. The total negative relevance of 30% seems excessive.

3.1.2. Sensitivity Analysis (SA)

On the other hand, Sensitivity Analysis ([47], [48]) explains a prediction based on the model's locally evaluated gradient. This method assumes that the most relevant input features are the ones that would cause the biggest change to the output if their values were changed, i.e. the features that the model is most sensitive to.

Mathematically, sensitivity is calculated by the Jacobian, i.e. for each input variable i, as the partial derivative of function $f(x)$ near i:

$$R_i = \left|\left| \frac{\partial}{\partial x_i} f(\mathbf{x}) \right|\right|. \tag{27}$$

If the local rate of change R_i has large magnitude it means that the output is sensitive to input i in the vicinity of x. This operation can be applied to any of the layers of a neural network, or one can use the chain rule to derive an expression for the Jacobian similar to how the gradient of the loss function is derived and calculate the sensitivity of the network output to the network input.

In Figure 6, we present example maps of pixels that have high contribution to the classification result of a convolutional DNN. The maps were extracted using a single back-propagation pass through the model.

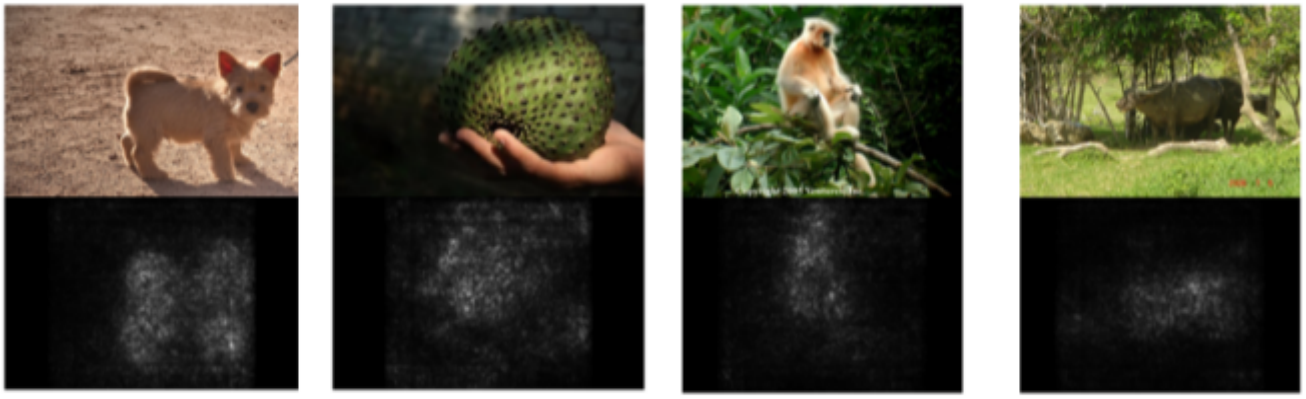

Figure 6 Example maps of pixel contribution in the classification process of convolutional DNN [47].

3.1.3. Local Interpretable Model-agnostic Explanations (LIME)

Ribeiro et.al. [49] present a system that treats a model as a black box but still manages to produce a valid explanation on which features have the greatest impact on the prediction. The main intuition of their system is that the explanation may be derived locally from records generated randomly in the neighborhood of the record to be explained, and weighted according to their proximity to it. They achieve this by using interpretable classifiers in a locally faithful manner. Even though an interpretable model may not be able to approximate the black box model globally, approximating it in the vicinity of an individual instance may be feasible.

Formally, for a model $f : \Re^d \rightarrow \Re$ and the original representation $x \in \Re^d$ of an instance being explained, model $g : \Re^{d'} \rightarrow \Re, g \in G$, is an interpretable model from a class of potentially interpretable models G, such that $g(x')$ approximates $f(x)$, where $x' \in \Re^{d'}$ is the interpretable representation of the instance.

$\Pi_x(z)$ is a proximity measure between an instance z to x, so as to define locality around x and $\Omega(g)$ is a measure of complexity of g.

Additionally, $\mathcal{L}(f, g, \Pi_x)$ is a measure of how unfaithful g is in approximating f in the locality defined by $\Pi_x(z)$.

In order for the explanation $\xi(x)$ to be both locally faithful and at the same time interpretable by humans, it is obtained by solving:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \Pi_x) + \Omega(g) \tag{28}$$

The primary intuition behind LIME, is presented in Figure 7, where instances are sampled both in the vicinity of x (which have a high weight due to Π_x) and far away from x (low weight from Π_x). Even though the original model may be too complex to explain globally (the blue/pink background line), LIME presents an explanation (linear in this case) that is locally faithful (separates correctly the crosses from the dots), where the locality is captured by Π_x.

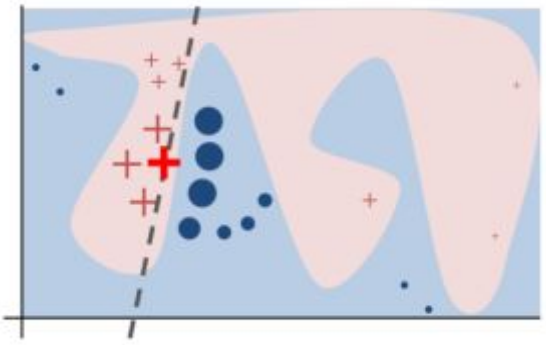

Figure 7 Toy example explaining local faithfulness of LIME explanations [49]

For image classification, one can use "super-pixels", computed using any standard algorithm. An interpretable representation of the image would be a binary vector where 1 indicates the original super-pixel and 0 indicates a grayed out super-pixel. This particular choice of Ω makes directly solving equation (28) intractable, but one can approximate it by first selecting K features with Lasso and then learning the weights via least squares. The choice of interpretable representations is crucial as to the type of contributing features that can be explained. For example, the presence or absence of super-pixels cannot explain *sepia* toned images being classified as *retro*.

As an example, in Figure 8 we present an example of explanations provided by LIME for the top 3 predicted classes of the Inception model when shown the image (a). The top predicted class "Electric Guitar" with a probability of 32% is explained in (b), the prediction of "Acoustic Guitar" with a probability of 24% is explained in (c) and finally in (d) is the explanation of "Labrador" with 21% probability. While the top prediction is wrong, since the correct type of guitar is acoustic, one can observe that the intuition of the model is correct to identify the neck of a guitar as being part of the "Electric Guitar" class.

Figure 8 Explanations produced by LIME for the top 3 predicted classes of an image [49]

3.2. METHODOLOGY

In the FANDANGO project a major goal is to provide a trustworthiness score for emerging news. In order to decide if a news article is trustworthy, the trustworthiness of each of the modalities of the article have to be examined. Then, a final aggregated trustworthiness score is provided, that takes into account the predictions for each of the modalities. A full description of the methods used for each modality is provided in the other deliverables of WP4.

The textual information that exists in an article is analyzed in two ways to provide semantic information about the content of the article in terms of named entities mentioned, location that the article refers to and time period that is analyzed, and to define a trustworthiness score using natural language processing methods. So far, the trustworthiness score is calculated by a linear model that uses syntactic rules as features. The linearity of the algorithm used means that the importance of each feature is easily tracked and reported in a form of weights, therefore explainable. Also, the syntactic rules are easily interpretable by the end-user.

An other source of trustworthiness for the article comes from the gathered metadata. The metadata include the authors that have written the article and the publisher that distributed the article. The credibility of both publishers and authors is calculated by measuring the importance in the context of the FANDANGO project and the number and trustworthiness of their previous articles. The score is calculated using statistical models which are both explainable and interpretable.

In parallel, semantic information about the spatial and temporal context of the article is also drawn from the media included. In order to decide the trustworthiness of a media object, the fact that it has not been manipulated in any way has to be discerned. Instead of using a single classifier on the input media to discern if it is manipulated, an ensemble of deep learning models is used to detect various forms of manipulation. Then, the predictions are concatenated to form the input of a classifier that predicts if the media has been manipulated.

For the images, the output of each of the models is a segmentation mask, i.e. an image of the same size where each pixel has a probability score that the pixel has been manipulated by the specific attack (Figure 9). The outputs are subsequently concatenated to the original image in the channel axis to create a multichannel image.

The classifier learns to predict if an image has been manipulated based on the probability masks. One advantage is that the classifier can gauge the efficacy of each ensemble model and learn to pay attention to specific models or ignore ineffectual models. The second advantage is that latent patterns from the ensemble models can be identified and taken advantage of to improve the classification score.

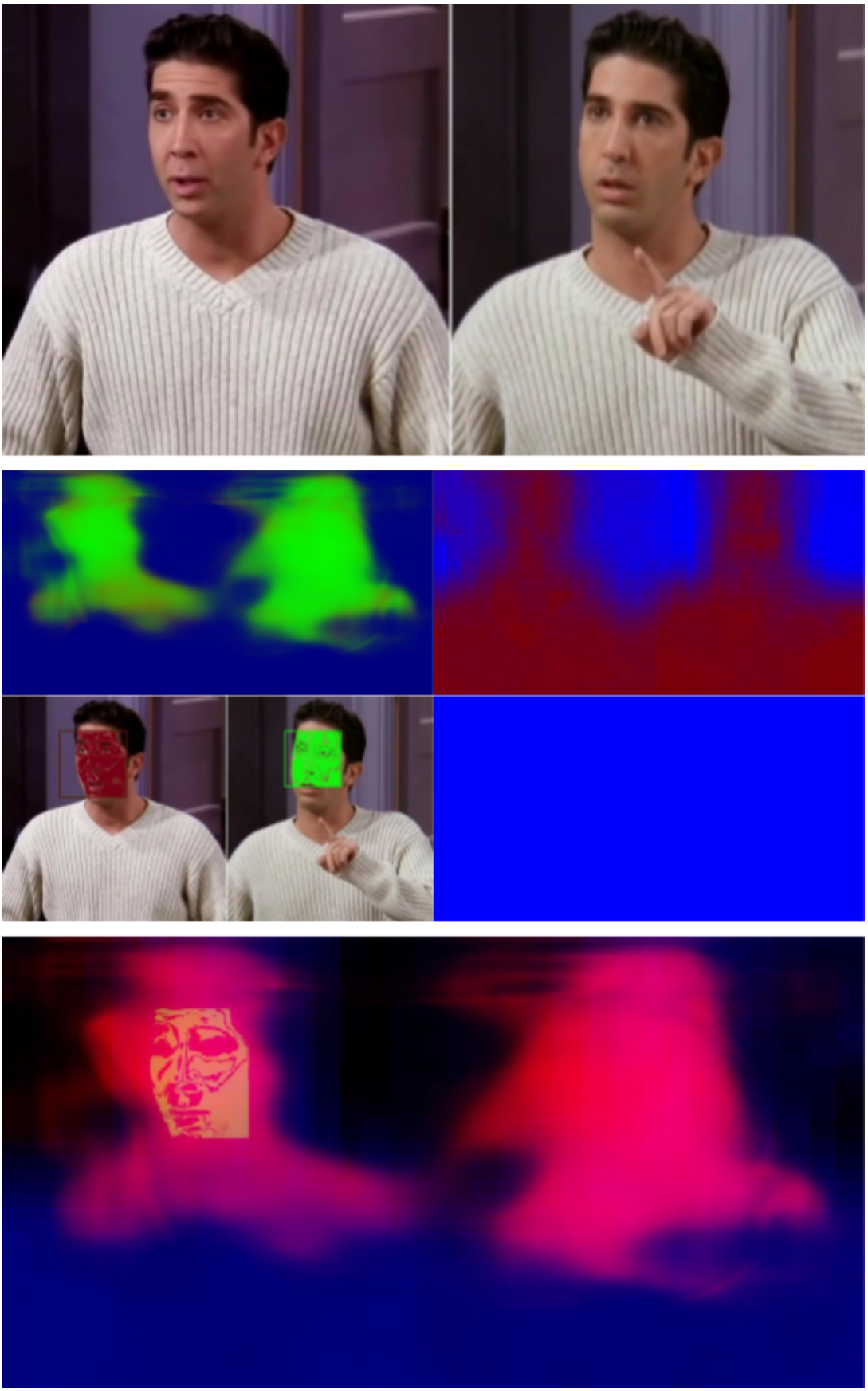

Figure 9 Example input of individual predictions to the image ensemble classifier

Finally, after all the scores for the various modalities have been calculated, a trustworthiness score for the article needs to be calculated. The fusion classifier will be trained to form one final aggregation score based on the trustworthiness scores of the different media modalities. It is imperative that this score is very well understood by the end users and how the classifier came up with it because this score can help the end user reach a final decision.

3.3. Implementation

The FANDANGO project will provide the end users with explanations on all the scores that are presented to the user. The scores that are provided for the textual and metadata modalities are from algorithms that can be explained and interpreted by the user.

3.3.1. Explaining the Media Score

The visual modality analysis provides, besides the score, segmentation masks from the selected manipulation detectors. While the examination of the masks can provide an explanation on what the different models have detected, the final classification score might not be reflected properly. Therefore an explanation will also be provided for the single trustworthiness score of the media.

For detecting if an image is manipulated or not, six independent models are used. The BusterNet [50] model and our extension, the Quadratic model, are used to detect copy-move attacks. The MantraNet [51], EXIF [52] and MaskRCNN [53] models are used as general purpose manipulation detectors. A detection of deep fakes, targeted on the faces of people present in the image is performed by the FaceForensics [54] model. All these models were trained on their own in various public [55] datasets. Their performance is measured taking into account the correct number of pixels that were correctly identified as being manipulated.

As a first step, a shallow VGG style convolutional network will be used for the classification process. This model can leverage the segmentation masks and extract features in the neighbourhood of each pixel. We deem that 3 convolution layers of 256 neurons size and a ReLU activation can extract enough information to properly aggregate the manipulation probabilities. The convolution layers will be followed by 2 fully connected layers, using the sigmoid activation, of sizes 256 and 64 respectively.

Even though it is common practice to start with pre-trained networks, the nature of the input for this model is very different from an image, making any pre-training void. Therefore the model will be trained from scratch using a concatenation of the training datasets used in the aforementioned models but with a binary classification target.

In order to provide an explanation of the classification, we use the LIME explanatory model. LIME provides a masked copy of the input hyperimage that has the most relevance to the output decision. Because an image has only 3 channels, in order to present the end-user with an interpretable explanation we juxtapose the explanation mask to the original image.

The module will be implemented as a RESTful service that accepts a JSON object with the analysis results of the 6 segmentation models, the input to the model, and the classification result. After calculating the explanation, the service will return a JSON object with the explanation mask, in the form of a url to a stored image that can be shown to the end-user.

3.3.2. Explaining the Fusion Score

Regarding the textual information, the corresponding input to the fusion classifier will be the trustworthiness score for each one of the text modality samples. In a similar way the classifier will be fed with credibility/trustworthiness scores calculated based on each article's metadata and visual content.

Before the learning process begins, all the scores will be normalized in the context of the preprocessing step of the proposed method. Then, the extracted features from all the article modalities will be mapped to one single space, belonging to one of the existing media modalities (text, metadata, visual content).

In the next step all the individual scores will be mapped to multiple representations in the new space, by initializing the SVD transformation matrix. This way the distance between the samples of each modality will be calculated., i.e., the distance between text, metadata and visual trustworthiness scores will be measured, to determine their relevance. This step ensures that the distance between the data of the same class will be small, while the distance between data of different classes will be large and the obtained mappings are used for measuring the distance in the means of relevance, between the text, metadata and visual info trustworthiness.

The transformed scores will finally be fed to the graph, in order for the distances that have been measured to be kept. The resulting representation will be acquired in a new space where the binary classification will be performed, to return an aggregated trustworthiness score. This score will carry information about the credibility of each modality and as a result, will be a truly meaningful indicator about article's trustworthiness.

For explaining the fused score decision, we use the LIME explanatory method. LIME will provide a relevance value for the contribution of each of the modalities in the final decision. For example a response would be that an article is 70% trustworthy and that the most contributing factor for this score, by 42%, is that the text modality is deemed 57% trustworthy and by 26% contribution the fact that the media modality is deemed as 81% trustworthy.

4. CONCLUSIONS

The FANDANGO platform aims to provide the tools necessary for an end-user to discern if a news item is valid or not. In order for a news item to be valid, all of the different components of the article (text, media etc.) must be authentic, unaltered and consistent in semantic, spatial and temporal level.

The combination of information from various parts of an article, into a single decision making function, can be as simple as calculating an average. Nonetheless, such a simple method could never hold up to a real world case. Moreover, some of the clues, regarding the trustworthiness of a news item, can only be found by examining multiple modalities simultaneously. Therefore, a proper machine learning classification model is imperative for an informed decision to be made.

Current state-of-the-art in fusion and multi-modal classification models, presents a multitude of approaches, each with strengths and weaknesses. A novel methodology is proposed, solving a number of the shortcomings of other methods, allowing for more accurate results. The validity of this method is tested on 3 distinct public datasets. The proposed methodology is implemented in the FANDANGO platform as a service, offering the end-user a trustworthiness measure that takes into account all the semantic, spatial and temporal information that can be acquired from an article.

It is a well known drawback of machine learning algorithms that the internal processes used to reach a conclusion are hard to be interpreted by humans [43]. Nonetheless, one of the most important factors for an end-user to trust machine learning decisions is to be able to understand the reason that such a decision was made.

In the FANDANGO platform, an interpretable explanation is provided in each decision making process, i.e. for every classifier. While the explanations provided are easily understood, a further step forwards would be to translate the explanations to plain text. So, instead of returning a list of positively and negatively contributing words or a percentage of relevance for each modality, a natural language text could be provided by the explanation method. The most challenging modality to interpret in natural language is the media one, where a combination of image detection, image segmentation and image captioning methods has to be achieved and all three disciplines are still major research topics [56].

REFERENCES

[1] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: An overview of methods, challenges, and prospects, Proceedings of the IEEE 103 (2015) 1449–1477.

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696.

[3] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (2013) 28 – 44.

[4] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, arXiv preprint arXiv:1705.09406 (2017).

[5] A. Shahroudy, G. Wang, T.-T. Ng, Multi-modal feature fusion for action recognition in rgb-d sequences, in: Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on, IEEE, pp. 1–4.

[6] H. LI, J. Sun, X. Zongben, L. Chen, Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network, IEEE Transactions on Multimedia (2017).

[7] Z. Liu, L. Zhang, Q. Liu, Y. Yin, L. Cheng, R. Zimmermann, Fusion of magnetic and visual sensors for indoor localization: Infrastructure-free and more effective, IEEE Transactions on Multimedia 19 (2017) 874–888.

[8] M. Guillaumin, J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, pp. 902–909.

[9] J. Geng, Z. Miao, X.-P. Zhang, Efficient heuristic methods for multimodal fusion and concept fusion in video concept detection, IEEE Transactions on Multimedia 17 (2015) 498–511.

[10] F. Sun, D. Harwath, J. Glass, Look, listen, and decode: Multimodal speech recognition with images, in: Spoken Language Technology Workshop (SLT), 2016 IEEE, IEEE, pp. 573–578.

[11] S. Receveur, D. Scheler, T. Fingscheidt, A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition, in: Situated Dialog in Speech-Based Human-Computer Interaction, Springer, 2016, pp. 179–192.

[12] D. Rafailidis, S. Manolopoulou, P. Daras, A unified framework for multimodal retrieval, Pattern Recognition 46 (2013) 3358–3370. 31

[13] S. Thermos, G. T. Papadopoulos, P. Daras, G. Potamianos, Deep affordance-grounded sensorimotor object recognition, margin 17 (2017) 35.

[14] F. Pala, R. Satta, G. Fumera, F. Roli, Multimodal person reidentification using rgb-d cameras, IEEE Transactions on Circuits and Systems for Video Technology 26 (2016) 788–799.

[15] F. Destelle, A. Ahmadi, N. E. O'Connor, K. Moran, A. Chatzitofis, D. Zarpalas, P. Daras, Low-cost accurate skeleton tracking based on fusion of kinect and wearable inertial sensors, in: Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, IEEE, pp. 371–375.

[16] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[17] L. Xu, X. Wu, K. Chen, L. Yao, Multi-modality sparse representation-based classification for alzheimer's disease and mild cognitive impairment, Computer methods and programs in biomedicine 122 (2015) 182–190.

[18] E. A. Bernal, X. Yang, Q. Li, J. Kumar, S. Madhvanath, P. Ramesh, R. Bala, Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors, IEEE Transactions on Multimedia (2017).

[19] C. Tan, F. Sun, W. Zhang, J. Chen, C. Liu, Multimodal classification with deep convolutional-recurrent neural networks for electroencephalography, in: International Conference on Neural Information Processing, Springer, pp. 767–776.

[20] K. Kalimeri, C. Saitis, Exploring multimodal biosignal features for stress detection during indoor mobility, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, pp. 53–60.

[21] R. Wagh, S. Darokar, S. Khobragade, Multimodal biometrics features with fusion level encryption, International Journal of Engineering Science 5246 (2017).

[22] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, A. D. N. Initiative, et al., Multi-modal classification of alzheimer's disease using nonlinear graph fusion, Pattern Recognition 63 (2017) 171–181.

[23] A. Patwardhan, G. Knapp, Aggressive actions and anger detection from multiple modalities using kinect, arXiv preprint arXiv:1607.01076 (2016).

[24] F. Cricri, M. J. Roininen, J. Leppanen, S. Mate, I. D. Curcio, S. Uhlmann, M. Gabbouj, Sport type classification of mobile videos, IEEE Transactions on Multimedia 16 (2014) 917–932.

[25] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard, Multimodal deep learning for robust rgb-d object recognition, in: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, IEEE, pp. 681–687.

[26] S. S. Mukherjee, N. M. Robertson, Deep head pose: Gaze-direction estimation in multimodal video, IEEE Transactions on Multimedia 17 (2015) 2094–2107.32

[27] F. Gürpinar, H. Kaya, A. A. Salah, Multimodal fusion of audio, scene, and face features for first impression estimation, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, pp. 43–48.

[28] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, T.-S. Chua, Multi-label visual classification with label exclusive context, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, pp. 834–841.

[29] P. Xie, E. P. Xing, Multi-modal distance metric learning, in: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press, pp. 1806–1812.

[30] M. Zeppelzauer, D. Schopfhauser, Multimodal classification of events in social media, Image and Vision Computing 53 (2016) 45–56.

[31] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, H. Sawhney, Multimodal fusion using dynamic hybrid models, in: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, IEEE, pp. 556–563.

[32] K. Sohn, W. Shang, H. Lee, Improved multimodal deep learning with variation of information, in: Advances in Neural Information Processing Systems, pp. 2141–2149.

[33] J. Li, Y. Wu, J. Zhao, K. Lu, Multi-manifold sparse graph embedding for multi-modal image classification, Neurocomputing 173 (2016) 501–510.

[34] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE transactions on pattern analysis and machine intelligence 37 (2015) 2085–2098.

[35] Z. Li, J. Tang, Weakly supervised deep metric learning for community-contributed image retrieval, IEEE Transactions on Multimedia 17 (2015) 1989–1999.

[36] J. Li, Zechao Tang, Weakly supervised deep matrix factorization for social image understanding, IEEE Transactions on Image Processing 26 (2017) 276–288.

[37] P. C. Hansen, The truncated svd as a method for regularization, BIT Numerical Mathematics 27 (1987) 534–553.

[38] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM international conference on image and video retrieval, ACM, p. 48.

[39] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019.

[40] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 198–213.

[41] D. Hu, X. Li, et. al., Temporal multimodal learning in audiovisual speech recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3574–3582.

[42] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 93.

[43] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE, 10(7):e0130140, 2015.

[44] L. Arras, G. Montavon, K.-R. Müller, and W. Samek. Explaining recurrent neural network predictions in sentiment analysis. In Proceedings of the EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA), pages 1–10, 2017.

[45] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2912–2920, 2016.

[46] G. Montavon, S. Bach, A. Binder, W. Samek, and K.R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65:211–222, 2017.

[47] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. Journal of Machine Learning Research, 11:1803–1831, 2010.

[48] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1135–1144.

[50] Wu, Y., Abd-Almageed, W., & Natarajan, P. (2018). BusterNet: Detecting copy-move image forgery with source/target localization. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 168-184).

[51] Wu, Y., Abd-Almageed, W., & Natarajan, P. (2019). ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9543-9552).

[52] Huh, M., Liu, A., Owens, A., & Efros, A. A. (2018). Fighting fake news: Image splice detection via learned self-consistency. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 101-117).

[53] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[54] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: *arXiv preprint arXiv:1901.08971*, (2019).

[55] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179.

[56] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4651-4659).