HAL
open science

CentralNet: a Multilayer Approach for Multimodal Fusion

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, Frédéric Jurie

To cite this version:

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, Frédéric Jurie. CentralNet: a Multilayer Approach for Multimodal Fusion. European Conference on Computer Vision Workshops: Multimodal Learning and Applications, Sep 2018, Munich, Germany. hal-01858560

HAL Id: hal-01858560

https://hal.science/hal-01858560v1

Submitted on 21 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CentralNet: a Multilayer Approach for Multimodal Fusion

Valentin Vielzeuf[1,2], Alexis Lechervy[2], Stéphane Pateux[1], and Frédéric Jurie[2]

[1] Orange Labs
[2] Normandie Univ., UNICAEN, ENSICAEN, CNRS
{valentin.vielzeuf,stephane.pateux}@orange.com
{alexis.lechervy,frederic.jurie}@unicaen.fr

Abstract. This paper proposes a novel multimodal fusion approach, aiming to produce best possible decisions by integrating information coming from multiple media. While most of the past multimodal approaches either work by projecting the features of different modalities into the same space, or by coordinating the representations of each modality through the use of constraints, our approach borrows from both visions. More specifically, assuming each modality can be processed by a separated deep convolutional network, allowing to take decisions independently from each modality, we introduce a central network linking the modality specific networks. This central network not only provides a common feature embedding but also regularizes the modality specific networks through the use of multi-task learning. The proposed approach is validated on 4 different computer vision tasks on which it consistently improves the accuracy of existing multimodal fusion approaches.

Keywords: Multimodal Fusion · Neural Networks · Representation Learning · Multi-task Learning

1 Introduction and Related Work

Multimodal approaches are key elements for many computer vision applications, from video analysis to medical imaging, through natural language processing and image analysis. The main motivation for such approaches is to extract and combine relevant information from the different modalities and hence take better decisions than using only one. The recent literature abounds with examples in different domains such as video classification [1, 2], emotion recognition [3–5], human activity recognition [6], or more recently food classification from pictures and recipes [7].

The literature on multimodal fusion [8–10] usually distinguishes the methods accordingly with the level at which the fusion is done (typically early vs late fusion). There is no consensus on which level is the best, as it is task dependent. For instance, Simonyan *et al.* [6] propose a two stream convolutional neural network for human activity recognition, fusing the modalities at prediction level. Similarly, for audiovisual emotion recognition, several authors report

better performance with late fusion approaches [11, 12]. In contrast, Arevalo *et al.* [13] propose an original Gated Multimodal Unit to weight the modalities depending on the input and achieve state of the art results on a textual-visual dataset, while Chen *et al.* [14] follow an early fusion hard-gated approach for textual-visual sentiment analysis.

Opposing early and late fusion is certainly too limited a view on the problem. As an illustration, Neverova *et al.* [15] applies a heuristic consisting in fusing similar modalities earlier than the others. Several hybrid or multilayer approaches have also been proposed, such as the approach of Yang *et al.* [16] doing fusion by boosting across all layers on human activity videos. Cătălina Cangea *et al.* [17] propose a multilayer cross connection from 2D to 1D to share information between modalities of different dimensions. A multilayer method is also applied on text and image multimodal datasets in the paper of Gu *et al.* [18]. Kang *et al.* [19] use a multilayer approach, aggregating several layers of representation into a contextual representation. These hybrid methods can be viewed as learning a *joint representation*, following the classification made by Baltruvsaitis *et al.* [20]. With this type of approach the different modalities are projected into the same multimodal space, e.g. using concatenation, element-wise products, etc.

Baltruvsaitis *et al.* [20] oppose *joint representations* with *coordinated representations* where some constraints between the modalities force the representations to be more complementary. These constraints can aim at maximizing the correlation between the multimodal representations, as in Andrew *et al.* [9] who propose a deep Canonical Correlation Analysis method. On their side, Chandar *et al.* [21] propose CorrNet using autoencoders. Neverova *et al.* [22, 23] propose modDrop and modout regularization, consisting in dropping modalities during the training phase. Finally, Hu *et al.* [5] applies an ensemble-like method to solve the problem of multimodal fusion for emotion classification.

This paper borrows from both visions, namely the joint representations and the coordinated representations. Our fusion method builds on existing deep convolutional neural networks designed to process each modality independently. We suggest to connect these networks using an additional central network dedicated to the projection of the features coming from different modalities into the same common space. In addition, the global loss allows to back propagate some global constraints on each modality, coordinating their representations. As an interesting property, the proposed approach automatically identifies which are the best levels for fusing the information and how these levels should be combined. The approach is multitask in the sense that it simultaneously tries to satisfy per modality losses as well as the global loss defined on the joint space.

The rest of the paper is organized as follows: the next section presents our contribution while Section 3 gives an experimental validation of the approach.

2 CentralNet

We refer to *multimodal fusion* as the combination of information provided by different media, under the form of their associated features or the intermediate

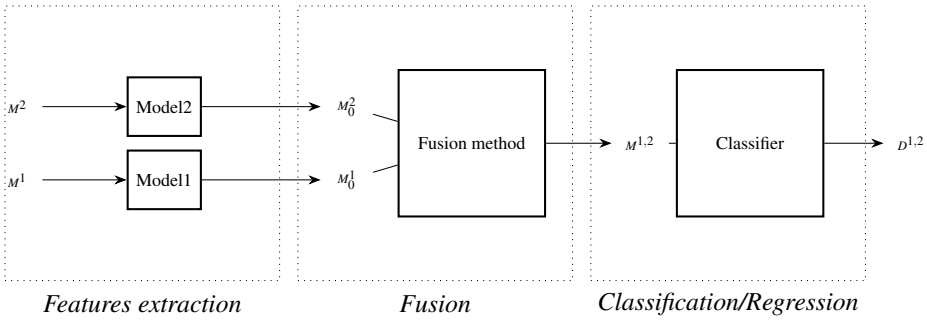

Fig. 1. Generic representation of a multimodal fusion model. M^1 and M^2 respectively denote modality 1 and modality 2, M_0^1 and M_0^2 are the modality features fed to the fusion method, $M^{1,2}$ is the joint representation produced by the fusion method, and $D^{1,2}$ the decision obtained from the joint representation.

decisions. More formally, if M^1 and M^2 denote the two media and D^1 and D^2 the decisions inferred respectively by M^1 and M^2, the goal is to make a better prediction $D^{1,2}$ using both M^1 and M^2. More than 2 modalities can be used. This paper addresses the case of classification tasks, but any other task, *e.g.* regression, can be addressed in the same way.

This paper focuses on the case of neural nets, for which the data are sequentially processed by a succession of layers. We assume having one neural net per modality, capable of inferring a decision from each modality taken in isolation, and want to combine them. One recurrent question with multimodal fusion is where the fusion has to be done: close to the data (early fusion), at the decision level (late fusion) or in between. In case of neural networks, the fusion can be done at any level between the input and the output of the different unimodal networks. For the sake of presentation, let us consider that the neural networks are split into 3 parts: the layers before the fusion (considered as being the feature generation part of the networks), the layers used for the fusion and finally the classification parts of the networks. This is illustrated by Figure 1.

For simplicity, we assume that the extracted features (at the input of the fusion layers) have the same dimensionality. If it is not the case, the features can be projected, *e.g.* with 1x1 convolutional layers or zero padded to give them the same size. In practice, the last convolution layers or the first dense layers of separately trained unimodal networks can be used as features.

2.1 CentralNet Architecture

The CentralNet architecture is a neural network which combines the features issued from different modalities, by taking, as input of each one of its layers, a weighted sum of the layers of the corresponding unimodal networks and of its own previous layers. This is illustrated in Figure 2(b). Such fusion layers can be

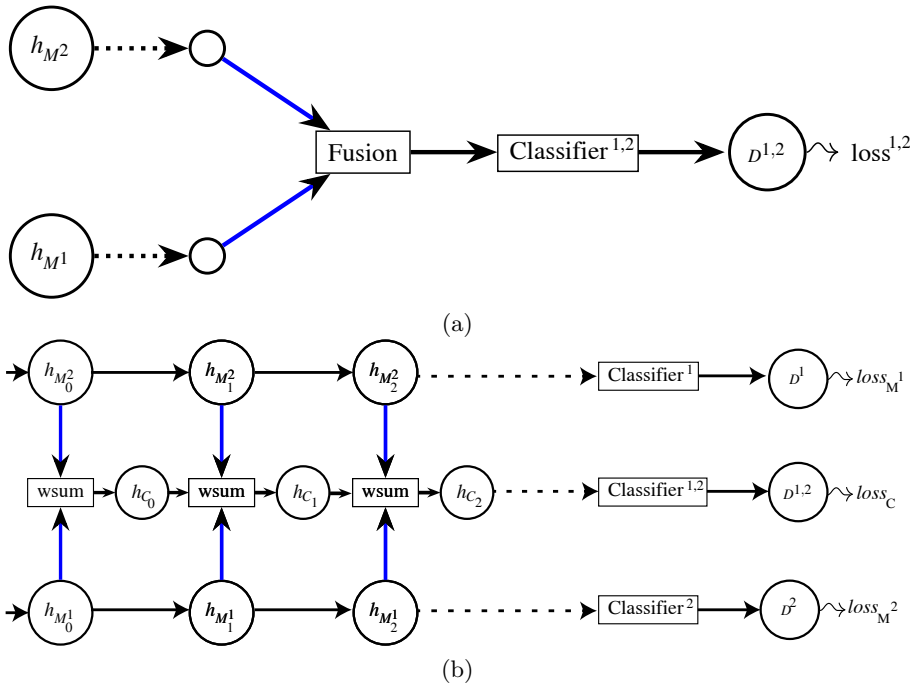

Fig. 2. (a) Basic fusion method, fusing the hidden representations of the modalities at a given layer and then using only joint representation. Fusing at a low-level layer is called early fusion while fusing at the last layer is called late fusion. **(b) Our CentralNet fusion model**, using both unimodal hidden representations and a central joint representation at each layer. The fusion of the unimodal representations is done here using a learned weighted sum. For the sake of simplicity, only the overall synoptic views of the architectures are represented. More details are provided in Section 2.

defined by the following equation:

$$h_{C_{i+1}} = \alpha_{C_i} h_{C_i} + \sum_{k=1}^{n} \alpha_{M_i^k} h_{M_i^k} \tag{1}$$

where n is the number of modalities, α are scalar trainable weights, $h_{M_i^k}$ is the hidden representation of each modality at layer i, and h_{C_i} is the central hidden representation. The resulting representation $h_{C_{i+1}}$ is then fed to an operating layer cell (which can be a convolutional or a dense layer followed by an activation function).

Regarding the first layer of the central network ($i = 0$), as we do not have any previous central hidden representation, we only weight and sum the representations of M^1 and M^2, issued from unimodal networks. At the output level, the last weighted sum is done between the unimodal predictions and the central prediction. Then, the output of the central net (classification layer) is used as the final prediction.

2.2 Learning the CentralNet model

All trainable weights of the unimodal networks, the ones of the CentralNet and the fusion parameters $\alpha_{M_i^k}$, are optimized together by applying a stochastic gradient descent using the Adam approach. The global loss is defined as:

$$loss = loss_C + \sum_k \beta_k loss_{M^k} \tag{2}$$

where $loss_C$ is the (classification) loss computed from the output of the central model and $loss_{M^k}$ the (classification) loss when using only modality k. The weights β_k are cross validated (in practice, $\beta_k = 1$ in all of our experiments).

As already observed by Neverova *et al.* [22], when dealing with multimodal fusion it is crucial to maintain the performance of the unimodal neural networks. It is the reason why the global loss includes the unimodal losses. It helps generalizations by acting as a multitask regularization. We name this method "Multi-Task" in the rest of the paper.

2.3 Implementation details

The α_{C_i} weights are initialized following a uniform probability distribution. Before training, the weighted sum is therefore equivalent to a simple average.

During our experiments, we also found out that rewriting Eq. (1) as:

$$h_{C_{i+1}} = \alpha_{C_i} h_{C_i} + \alpha_{modalities} \sum_{k=1}^{n} \alpha_{M_i^k} h_{M_i^k} \tag{3}$$

leads to better and stable performance.

On overall, CentralNet is easy to implement and can build on the top of existing architectures already known to be efficient for each modality. The number of trainable parameters dedicated to the fusion is less important than in other previous multilayer attempts such as [18], which may help to prevent overfitting. And even if the weighted sum is a simple linear operation, the network has the ability to learn complex joint representation, because of the non-linearity introduced by the central network.

Finally, the resulting values of the α allow some interesting interpretations on where the modalities are combined. For instance, getting $\alpha_{M_i^k}$ values close to 0 for $k > 0$ is equivalent to early fusion, while having all the α_{C_i} close to 0 up to the last weighted sum would be equivalent to late fusion.

3 Experiments

The proposed method is experimentally validated on 4 different multimodal datasets, namely Multimodal MNIST (a toy dataset), Audiovisual MNIST, Montalbano [24] and MM-IMDb [13], receiving each a separate section in the following. Each dataset is processed with a dedicated features extractor on which we

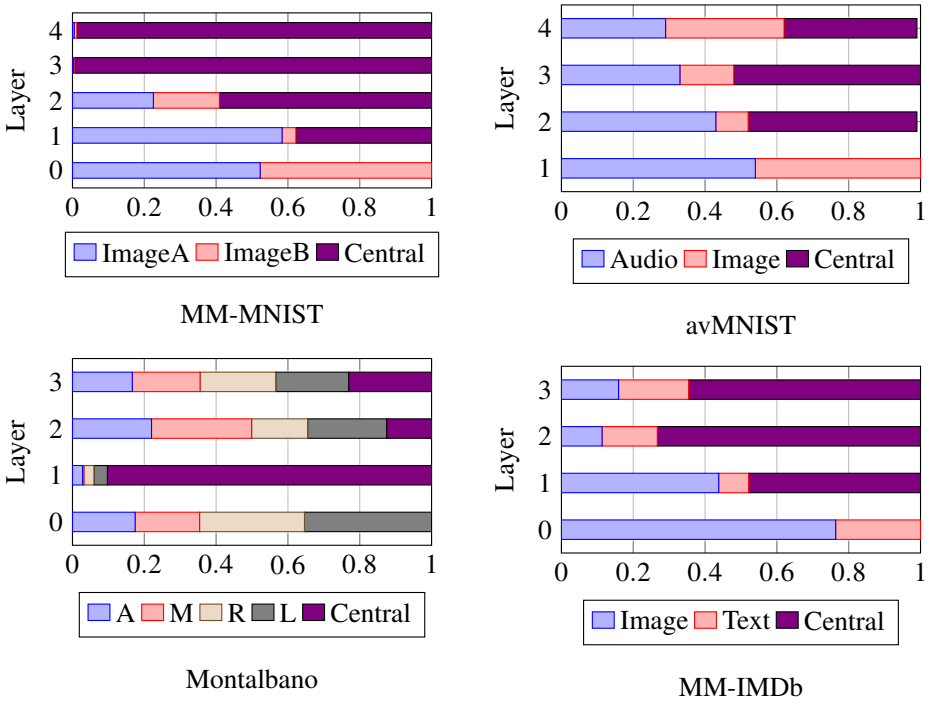

Fig. 3. Visualization of the α_i weights after training. They are displayed as the percentage given to each modality and to the central hidden representations across layers and datasets. We observe that the learned fusion strategy is different for each dataset.

plug our fusion method. Regarding the fusion networks, they are made of convolution+pooling or dense layers, with ReLU and batch normalization. The kernel size of the convolution is always 5x5 and the pooling stride is 2.

The performance of the proposed method is compared to 5 different fusion approaches, ranging from the simplest baseline to recent state-of-the-art approaches. (a) 'Weighted mean' is the weighted average of single modality scores. The weights are considered as some parameters of the model, learnt with the rest of the model. (b) 'Concat' consists in concatenating the unimodal scores and inferring the final score with a single-layer linear perceptron. (c) 'Concat+Multi-Task' is the same as 'Concat' but uses the same Multi-Task loss as with the CentralNet. (d) 'Moddrop' is implemented following Neverova *et al.* [22]. (e) The 'Gated Multimodal Unit' (GMU) is implemented following Arevalo *et al.* [13].

In the following, to assess the statistical significance of our results, the performance is averaged over 64 runs. The confidence interval at 99% is computed using the estimate of the standard deviation and the Student's law.

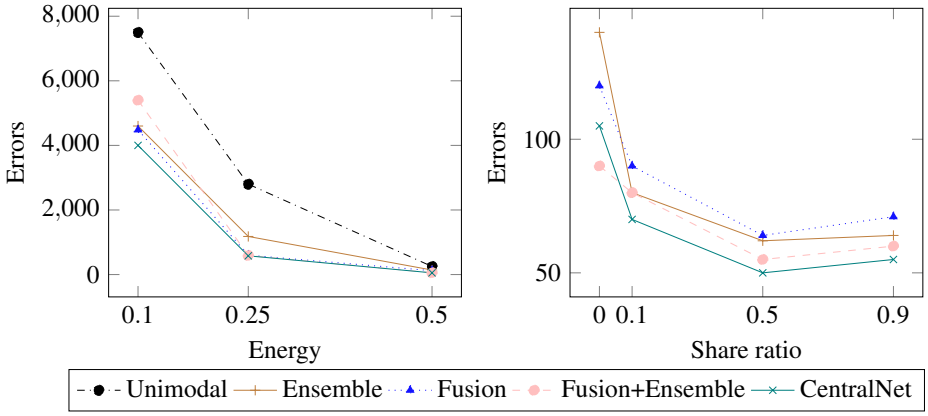

Fig. 4. Errors as a function of the energy per modality (left-hand side, share ratio=0.5) and of the share ratio (right-hand side, energy=0.5), for different fusion methods. Better viewed in color.

3.1 Multimodal MNIST

The 'Multimodal MNIST' dataset is a toy dataset made of pairs of images (A,B), computed from the MNIST dataset. A and B are supposed to be 2 views of the same MNIST image but from different (artificially generated) modalities. We produce them by computing a Principal Component Analysis of the original MNIST dataset. Each one of the 2 (artificial) modalities is created by associating with it a set of singular vectors. This allows to control the amount of energy provided to each modality, which is the sum of the *energy* contained in the chosen vectors, and the *share ratio*, defined here as the percentage of the singular vectors shared between modalities. Figure 5 shows some of these generated image pairs. The original MNIST contains 55000 training samples and 10000 test samples. We transformed all of these images into pairs of 28x28 images, following the process explained above.

Several authors, *e.g.*, [9, 22, 21, 23], generate a multimodal version of MNIST by dividing MNIST images into several smaller images (typically quarter of images) which are each considered as modalities. In contrast, our approach has the advantage of allowing to control two important factors: the amount of information per modality and the dependence between modalities.

The unimodal neural network architecture used with this dataset is the LeNet5 neural network [25]. It achieves 95 errors on the MNIST test set [25]. The architecture is composed of two convolutional layers, followed by two fully connected layers. In our version, batch normalization and dropout are added to further improve its performance. We measure the performance by counting how many of the 10000 images of the MNIST test set are misclassified. The CentralNet architecture in this case is therefore composed of three LeNet5, as described in Table1. The "Ensemble 3 classifiers" method also uses three LeNet5, while other methods are using two LeNet5, one for each modality. We use dropout (50%

Image A		Central		Image B	
Type	Size	Type	Size	Type	Size
Conv	14x14x32	Conv	14x14x32	Conv	14x14x32
Conv	7x7x64	Conv	7x7x64	Conv	7x7x64
Dense	1024	Dense	1024	Dense	1024
Pred	10	Pred	10	Pred	10

Table 1. The architecture of the CentralNet for the mMNIST dataset."Dense" layers are fully-connected layers followed by a ReLU activation, while "Pred" layers are fully-connected layers followed by softmax activation.

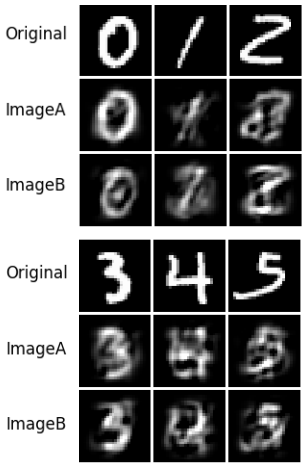

Fig. 5. MM-MNIST: some examples generated with half of the energy per modality and no sharing.

Method	Errors	Fusion Layer
Baseline	66 ±1.5	-
Ensemble 2 classifiers	64 ±1.3	-
Ensemble 3 classifiers	60 ±1.0	-
Fusion Subtract	**64 ±1.8**	**2**
Fusion Sum	68 ±2.1	2
Fusion Prod	71 ±2.1	2
Fusion+Ensemble Subtract	63 ±1.5	1
Fusion+Ensemble Sum	**56 ±2.1**	**0**
Fusion+Ensemble Prod	63 ±1.5	2
Baseline on one modality	230 ±2.7	-
Concat + Multi-Task	62 ±1.2	-
Moddrop[22]	60 ±1.5	-
Gated Multimodal Unit[13]	68 ±1.8	-
CentralNet	**53 ±1.2**	-

Table 2. Number of errors on the MM-MNIST test set for different methods, using 50% energy per modality and 50% of shared vectors.

dropping) on the fully connected layers and batch normalization. The learning rate is 0.01, the batch size is 128 and the model is trained on 100 epochs for all experiments, except for Moddrop and Gated Multimodal Units, where hyper-parameters are found by a random grid search. Thus for Moddrop, the learning rate is changed into 0.05 and the modality drop probability is of 0.2. For Gated Multimodal Units, the dropout is changed into 25% dropping.

First, we evaluate different alternatives for fusion (see Figure 2(b)) using element-wise sum, subtract and product, for several configurations of our toy dataset. The energy is in $\{0.1, 0.25, 0.5\}$ and share ratio in $\{0, 0.1, 0.5, 0.9\}$, allowing to assess the improvement given by fusion on each configuration. We also evaluate the Fusion+Ensemble method, *i.e.*, an ensemble of classifiers build on the top of the outputs of the fusion method (each modality make a prediction, as well as the fusion method, giving an ensemble of 3 classifiers). Finally, we also report the results of our CentralNet approach.

	Image		Central		Audio	
Part	Type	Output size	Type	Output size	Type	Output size
Features Extraction	Conv1	14 x 14 x 32			Conv1 Conv2 Conv3	56 x 56 x 8 28 x 28 x 16 14 x14 x 32
Fusion	Conv Dense Pred	7 x 7 x 64 1024 10	Conv Dense Pred	7 x 7 x 64 1024 10	Conv Dense Pred	7 x 7 x 64 1024 10

Table 3. The architecture of the CentralNet model on the avMNIST dataset.

We numbered the layers of LeNet5 from 0 (input level) to 4 (prediction level) and evaluate the methods for the 5 different fusion depth, in order to find out which one yields is the best. Figure 4 reports the performance of the different methods. The performance of the Fusion and Fusion+Ensemble methods are given in the case of their best fusion depth.

These results first underline the proportionality relation between the energy per modality and the error rate. It is also worth noting that not sharing enough or too much information between the modalities lowers the accuracy and the interest of a fusion approach. This observation is in line with [20, 8].

As shown in Table 2 the optimal fusion layer obtained for each method differs but is early. Other properties are highlighted: A complementarity between Fusion and Ensemble exists, as shown by the improvement brought by the Fusion+Ensemble method. Nevertheless, as soon as the modalities share a large amount of information, the Ensemble method outperforms the Fusion method. It implies that the benefit of the fusion depends on the nature of the dataset and can be null.

Independently to the chosen configuration, our CentralNet approach achieves the best results, except in the case of a null share ratio (first point of the right-hand side of the Figure 4). In this case, the modalities are not sharing information, so the better performance of the Fusion+Ensemble (fusing at layer 0) compared to CentralNet might be explained by the difficulty to find relation between independent modalities and thus constructing a stable joint representation from the learned weighted sum. A comparison with an Ensemble of 3 models applied on original images suggests that this performance does not come only from a larger number of parameters.

Table 3 shows that in the lowest layers of CentralNet, the modalities are taken into account, while on last layers the weight of central previous hidden layer dominates. This is in line with our observations on the Fusion+Ensemble results.

3.2 Audiovisual MNIST

Audiovisual MNIST is a novel dataset we created by assembling visual and audio features. The first modality, disturbed image, is made of the 28x28 PCA-projected MNIST images, generated as explained in the previous section, with

Method	Accuracy
Disturbed image	72.8 ±0.3
Audio	86.1 ±0.15
Weighted mean	94.7 ±0.12
Concat	93.7 ±0.17
Concat + Multi-Task	94.8 ±0.11
Moddrop[22]	94.8 ±0.10
Gated Multimodal Unit[13]	94.1 ±0.14
CentralNet	**95.0 ±0.12**

Table 4. Accuracy on the audiovisual MNIST dataset.

only 25% of the energy, to better assess the benefits of the fusion method. The second modality, audio, is made of audio samples on which we have computed 112x112 spectrograms. The audio samples are the pronounced digits of the Free Spoken Digits Database [26] augmented by adding randomly chosen 'noise' samples from the ESC-50 dataset [27], to reach the same number of examples as in MNIST (55000 training examples, 10000 testing examples).

For processing the image modality, we use the LeNet5 architecture [25], as in the previous section. For the audio modality, we use a 6-layer CNN, adding two convolution-pooling blocks. The whole architecture is detailed on Table 3.

We use dropout (50% dropping) on the fully connected layers and batch normalization. The learning rate is 0.001, the batch size is 128 and the model is trained on 100 epochs for all experiments, except Moddrop and Gated Multimodal Units, where hyper-parameters are found by a random grid search. Thus for Moddrop, the learning rate is changed into 0.005 and the modality drop probability is of 0.32. For Gated Multimodal Units, the dropout is changed into 35% dropping.

The performance is measured as the per sample accuracy on the 10000 test samples. We observe from Table 4, that the fusion methods are all performing better than unimodal ones. Both ensembles, Moddrop and simple weighted mean yield good performance but CentralNet performs best. Figure 3 shows that all the modalities are used at each layer, meaning that they all bring information.

3.3 Montalbano

The Montalbano dataset [24] gathers more than 14000 samples of 20 Italian sign gesture categories. These videos were recorded with a Kinect, capturing audio, skeleton joints, RGB and depth. The task is to recognize the gestures from the video data. The performance is measured as the macro accuracy, which is the average of the per class accuracy.

The features used in these experiments are those provided by Neverova *et al.* [22]: audio features (size 350), motion capture of the skeleton (size 350), RGB+depth left/right hands features (size 400). Features are zero-padded (if needed) to give vectors of size 400. The fusion architecture includes one multi-layer perceptron per modality, each having 3 layers of size: 400×128, 128×42,

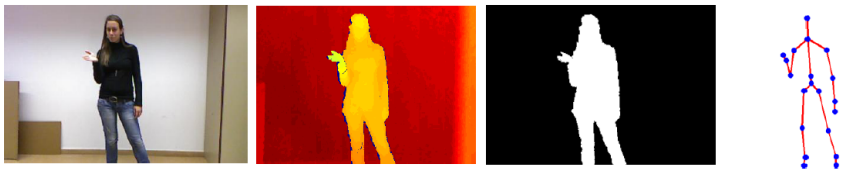

Fig. 6. The different raw visual modalities provided by the organizers of the ChaLearn challenge on Montalbano dataset. Neverova *et al.* [15] propose to focus on right and left hands, skeleton and audio.

Method	Accuracy
Left	46.0 ±0.7
Audio	59.3 ±0.3
Right	79.0 ±0.3
Mocap	88.0 ±0.3
Weighted mean	97.54 ±0.02
Concat	97.76 ±0.05
Concat + Multi-Task	98.02 ±0.04
Moddrop	98.19 ±0.03
Gated Multimodal Unit	97.98 ±0.04
CentralNet	**98.27 ±0.03**

Table 5. Accuracy on the Montalbano validation set (same protocol as [22]).

42×21. CentralNet architecture connects the 3 layers of the different modalities into a central network.

We use dropout (50% dropping) and batch normalization. The learning rate is 0.05 (we multiply the learning rate by 0.96 at each epoch), the batch size is 42 containing two samples of each class and the model is trained on 100 epochs for all experiments. For Moddrop, the modality drop probability is of 0.5.

Table 5 shows that the performance obtained with each modality varies from 46% (left hand) to 88% (mocap). Basic late fusion gives significant improvement, suggesting complementarity between modalities. CentralNet outperforms all other approaches. Figure 3 shows the weights of the different modalities at each level. At the first layer (layer 0), the weights reflect the dimensionality of the layers. At the next layer, almost no information is taken from the modalities, while at layers 2 and 3, the weight given to each modality and to the central representation are relatively similar. This may be interpreted as an hybrid fusion strategy, mixing "early" and "late" fusions.

3.4 MM-IMDb

The MM-IMDb dataset [13] comprises respectively 15552, 2608 and 7799 training, validation and test movies, along with their plot, poster, genres and other 50 additional metadata fields such as year, language, writer, director, aspect ratio, *etc.*. The task is to predict a movie genre based on its plot and on its poster (*cf.*

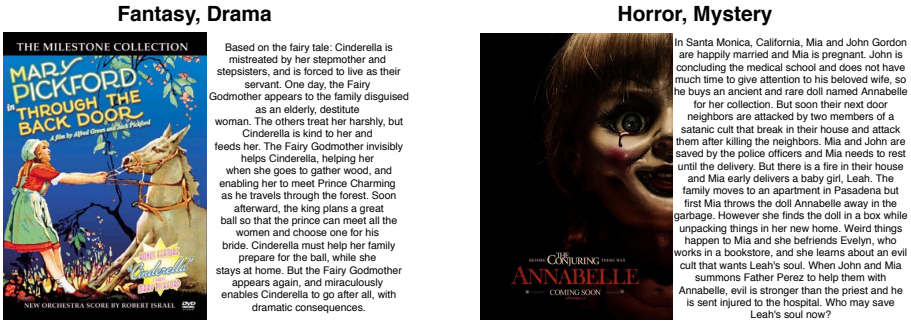

Fig. 7. Two movie samples extracted from the mm-IMDB dataset. For each, we can see the poster and the associated plot. The genres to predict are displayed on the top of the figure.

Text		Central		Visual	
Type	Size	Type	Size	Type	Size
Dense	2048	Dense	2048	Dense	2048
Dense	512	Dense	512	Dense	512
Pred	23	Pred	23	Pred	23

Table 6. Architecture of the CentralNet on the MM-IMDb dataset.

Figure 7). One movie can belong to more than one of the 23 possible genres. The task hence has to be evaluated as a multilabel classification task. As in [13, 7], we measure the performance with the micro, macro, weighted and per sample F1 scores. For these experiments, we use the features kindly provided by the authors [13]. The visual feature of size 4096 is extracted from the posters using the VGG-16 [28] network pretrained on Imagenet. The 300-d textual one are computed with a fine-tuned word2vec [29] encoder.

We build a multilayer perceptron on the top of the features of each modality. For both modalities, the network has 3 layers of size $input_size \times 4096$, 4096×512 and 512×23. The CentralNet architecture (see Table 6 is the same, taking 4096-d vectors as inputs, zero-padding the textual features to reach the visual features size.

We use dropout (50% dropping) and batch normalization. The learning rate is 0.01 and the batch size is 128. For Moddrop, the modality drop probability is of 0.25. The loss of the models is a cross entropy, but we put a weight of 2.0 on the positives terms to balance precision and recall. More formally, the loss is:

$$loss = -\log(2\sigma(pred))y - (1-y)log(1-\sigma(pred)) \tag{4}$$

with $\sigma(pred)$ the sigmoid activation of the last output of the network and y the multiclass label. As recommended by Arevalo *et al.* [13], we also use early stopping on the validation set.

Table 7 reports the performance measured during the different experiments. First of all, the worst confidence interval we observe is very small, of the order of

Method	Micro	Macro	Weighted	Samples
Text (alone)	0.602	0.489	0.585	0.606
Image (alone)	0.478	0.256	0.421	0.484
Weighted mean	0.635	0.550	0.626	0.634
Concat	0.611	0.506	0.599	0.614
Concat + Multi-Task	0.623	0.528	0.613	0.622
Moddrop[22]	0.624	0.526	0.614	0.625
Gated Multimodal Unit [13]	0.630	0.541	0.617	0.630
CentralNet	**0.639**	**0.561**	**0.631**	**0.639**

Table 7. F1 scores of the different methods on the MM-IMDb test set.

± 0.001. For making the table more readable, we do not include it. Second, one can observe that the textual modality clearly outperforms the visual one. Third, we note that even the basic fusion methods, such as the concatenation of the features, improve the score. Finally, the Concat+Multi-Task and Concat+ModDrop methods are outperformed by a significant margin by Gated Multimodal Unit and CentralNet, which is giving the best performance. Figure 3 shows that CentralNet gives more weight to the first layers, indicating that an "early fusion" strategy is privileged in this case, even if the two modalities contribute significantly at all levels.

4 Conclusions

This paper introduced a novel approach for the fusion of multimedia information. It consists in a joint representation having the form of a central network connecting the different layers of modality specific neural networks. The loss of this central network not only allows to learn how to combine the different modalities but also adds some constraints on the modality specific networks, enforcing their complementary aspects. This novel model achieves state-of-the-art results on several different multimodal problems. It also addresses elegantly the *late versus early* fusion paradigm.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. CoRR **abs/1609.08675** (2016)
2. Wang, Z., Kuan, K., Ravaut, M., Manek, G., Song, S., Yuan, F., Seokhwan, K., Chen, N., D'Haro-Enríquez, L.F., Luu, A.T., Zhu, H., Zeng, Z., Cheung, N., Piliouras, G., Lin, J., Chandrasekhar, V.: Truly multi-modal youtube-8m video classification with video, audio, and text. CoRR **abs/1706.05461** (2017)
3. Dhall, A., et al.: Collecting large, richly annotated facial-expression databases from movies. (2012)
4. Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Pantic, M.: Summary for avec 2017: Real-life depression and affect challenge and workshop. In: Proceedings of the 2017 ACM on Multimedia Conference, ACM (2017) 1963–1964

5. Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.: Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM (2017) 553–560
6. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. (2014) 568–576
7. Kiela, D., Grave, E., Joulin, A., Mikolov, T.: Efficient large-scale multi-modal classification. CoRR **abs/1802.02892** (2018)
8. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16**(6) (2010) 345–379
9. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning. (2013) 1247–1255
10. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE **103**(9) (2015) 1449–1477
11. Kim, D.H., Lee, M.K., Choi, D.Y., Song, B.C.: Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM (2017) 529–535
12. Vielzeuf, V., Pateux, S., Jurie, F.: Temporal multimodal fusion for video emotion classification in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM (2017) 569–576
13. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. In: ICLR Worshop. (2017)
14. Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM (2017) 163–171
15. Neverova, N., Wolf, C., Taylor, G.W., Nebout, F.: Multi-scale deep learning for gesture detection and localization. In: Workshop at the European conference on computer vision, Springer (2014) 474–490
16. Yang, X., Molchanov, P., Kautz, J.: Multilayer and multimodal fusion of deep neural networks for video classification. In: Proceedings of the 2016 ACM on Multimedia Conference, ACM (2016) 978–987
17. Cangea, C., Velickovic, P., Liò, P.: Xflow: 1d-2d cross-modal deep neural networks for audiovisual classification. CoRR **abs/1709.00572** (2017)
18. Gu, Z., Lang, B., Yue, T., Huang, L.: Learning joint multimodal representation based on multi-fusion deep neural networks. In: International Conference on Neural Information Processing, Springer (2017) 276–285
19. Kang, M., Ji, K., Leng, X., Lin, Z.: Contextual region-based convolutional neural network with multilayer fusion for sar ship detection. Remote Sensing **9**(8) (2017) 860
20. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
21. Chandar, S., Khapra, M.M., Larochelle, H., Ravindran, B.: Correlational neural networks. Neural computation **28**(2) (2016) 257–285
22. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8) (2016) 1692–1706

23. Li, F., Neverova, N., Wolf, C., Taylor, G.: Modout: Learning to fuse modalities via stochastic regularization. Journal of Computational Vision and Imaging Systems **2**(1) (2016)
24. Escalera, S., Baró, X., Gonzalez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: Workshop at the European Conference on Computer Vision, Springer (2014) 459–473
25. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
26. Jackson, Z.: free-spoken-digit-dataset. https://github.com/Jakobovski/decoupled-multimodal-learning (2017)
27. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia, ACM (2015) 1015–1018
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
29. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111–3119