Team name	UPC-STP (Team 7)
Team leader name	Xavier Giró-i-Nieto
Team leader address, phone number and email	xavier.giro@upc.edu Campus Nord UPC (Modul D5 Jordi Girona 1-3 08034 Barcelona (Catalonia)
Rest of team members	Amaia Salvador, Daniel Manchón-Vizuete, Matthias Zeppelzauer and Andrea Calafell
Team website URL (if any)	https://imatge.upc.edu/web/resources/cultural-event-recognition-computer-vision-software

Title of the contribution

Cultural Event Recognition by Fine-tunning a Convolutional Network and Temporal Refinement

General method description

We fine tuned the a pretrained Convolutional Network (CaffeNet) using *Caffe*, a deep learning framework, using at first only the training data (partitioning it as 80% for traininig and 20% for validation). Once the validation labels were provided, we fine tuned our network with the remaining 20% of training images using the real validation data.
The last layer of our fine tuned network gives us the confidence score for an image for each of the classes. Results using those scores improved the baseline, but still we tried some late fusion strategies training an SVM on the neural codes generated on each of the last three layers of the network (FC6, FC7 and FC8). We combined the descriptors extracted from both our fine tuned network and the pretrained one, achieving our maximum result by adding a final temporal refinement. The temporal refinement was applied only to images with time stamps in their EXIF metadata, where high classification scores based on visual features were penalized when their time stamp did not match well an event-specific temporal distribution learned from the training and validation data.

References

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Describe data preprocessing techniques applied (if any)	We did not preprocess data, but created an interactive visualization tool to look at it in a convenient way for all team members. The tool can be accessed from http://178.62.195.161:8080/assets/#/
Describe features used or data representation model (if any)	We used the FC6, FC7 and FC8 layers of *CaffeNet*, both for the network trained on ImageNet and the network fine-tuned with training data. These features were also complemented by a time stamp extracted from EXIF metadata, where available.
Dimensionality reduction technique applied (if any)	None

Segmentation strategy used (if any)	
Classifier or method used to train and validate your results (if any)	We used SVMs with linear kernel. A separate SVM was trained for each of the six FC layers (coming from the fine-tuned and the untuned ImageNet network). The output of the SVMs are probabilities for all 50 classes for each image. The probabilities of each SVM are weighted by temporal models learned from the training and validation data. In the next step the temporally refined SVM outputs are combined by decision fusion. For this purpose, the SVM outputs are concatenated, yielding a feature vector of 50x6 values per image. These feature vectors are input to a top-level SVM that fuses the individual lower-level predictions and generates the final scores for each image.
Large scale strategy (if any)	

Transfer learning strategy (if any)	We fine-tuned the CaffeNet previously trained with the ImageNet dataset with the training and validation data provided by the ChaLearn organizers. This way, the parameters of the network were not learned from scratch but just adapted to the Chalearn dataset
Compositional model used (scene context representation), i.e. pictorial structure (if any)	
Other technique/strategy used not included in previous items (if any)	At first, we used as a classifier the softmax one provided at the last layer of Caffe, which directly gave us the probability of the 50 classes for each image. However we obtained better results by discarding this classifier and training a linear or non-linear SVM from the fully connected layers. To improve the fine tuning we tried to use some extra images obtained of Flickr, but these images introduced some noise which decreased results, also after trying to filter them by user popularity, time stamp and location. Also the fine tuning of the *Caffe* was done with different epochs and with different modification of the weights of the layers.
Method complexity analysis	The presented method follows the basic architecture adopted in several state of the art approaches for image classification. It combines a visual classification with a temporal modelling, which provides two independent architectures which are fused at a certain stage.

Qualitative advantages of the proposed solution	The solution provides a fast response at test time as the feature extraction and model assessment is computationally light, especially if a GPU is available. This architecture may be easily ported to problems such as social event recognition in user generated content, or news event recognition from social media feeds.
Results of the comparison to other approaches (if any)	
Novelty degree of the solution and if it has been previously published	The solution follows the most common practices in the fine-tuning of convolutional networks. The main novelty is the addition of a temporal filtering, as well as combining the features from different layers of both fine-tuned and non-fine tuned networks. This work has not been previously published.

Language and implementation details (including platform, memory, parallelization requirements)	The source code has two main parts. The feature extraction form Caffe has been run from Python scripts, while the temporal modeling and the SVM classifier has been run from Matlab.
Human effort required for implementation, training and validation?	The project was mainly developed as a full time task for a Phd student and an associate professor during three weeks, in addition to a part-time support of a software engineer, a bachelor thesis student and another associate professor.
Training/testing expended time?	Fine-tuning of the CaffeNet network lasts around 4 hours, while feature extraction for the test data took around 15 minutes, in both cases, a on a high-end GPU NVidia Tesla K20m. SVM training and testing requires between 10-20 minutes.
General comments and impressions of the challenge	The challenge is interesting and visually sounding. The dispersion of the data sources (Chalearn site, CodaLab, submission site) has made the task a bit too much difficult to manage. Also the existance of two help forums (codalab and chalearn Google Groups) introduced some confusion. Organizers have been responsive to our questions.