# XFlow: 1D ⇝ 2D Cross-modal Deep Neural Networks for Audiovisual Classification

Cătălina Cangea, Petar Veličković and Pietro Liò
University of Cambridge
{catalina.cangea, petar.velickovic, pietro.lio}@cl.cam.ac.uk

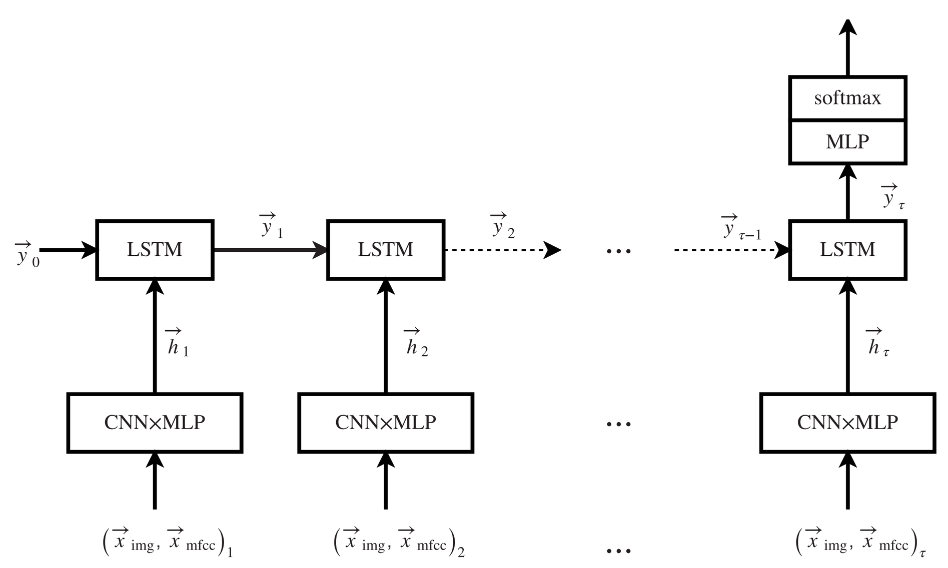
UNIVERSITY OF CAMBRIDGE

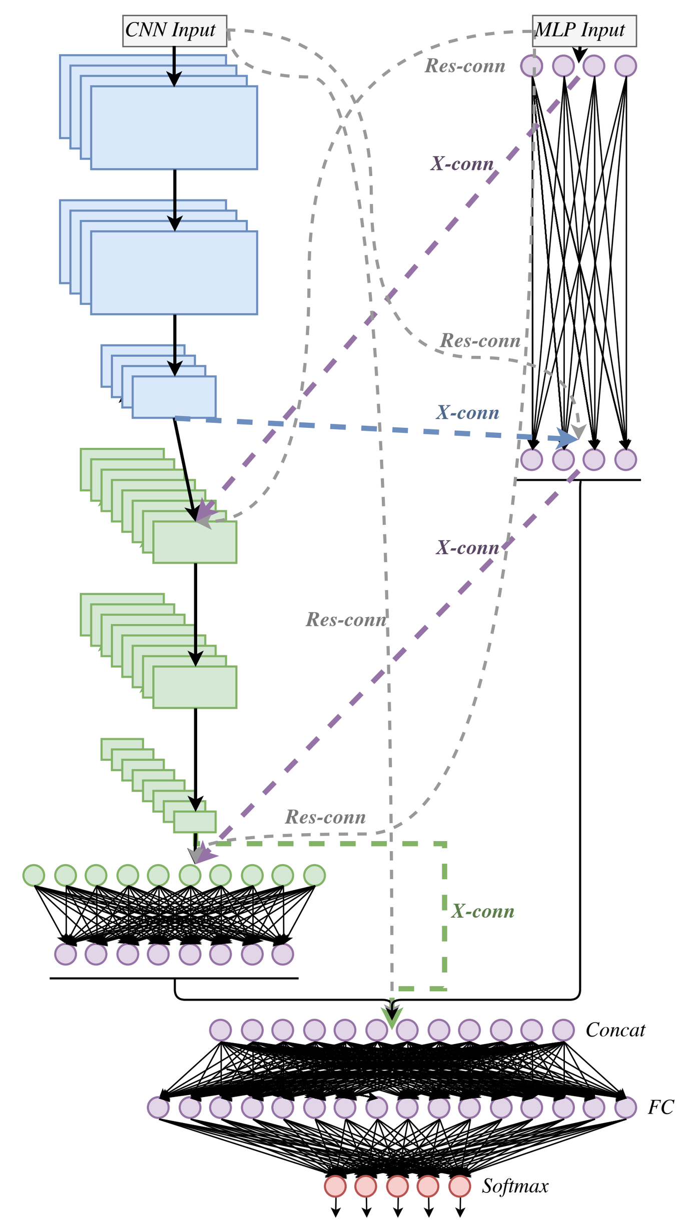
## Abstract

We propose two *multimodal deep learning architectures* [1] that allow for cross-modal dataflow (XFlow) between the feature extractors, thereby extracting more interpretable features and obtaining a better representation than through unimodal learning, for the same amount of training data. These models can usefully exploit correlations between audio and visual data, which have a different dimensionality and are therefore *nontrivially exchangeable*. Our work improves on existing multimodal deep learning methodologies in two essential ways: (1) it presents a novel method for performing cross-modality (*before* features are learned from individual modalities) and (2) extends the previously proposed *cross-connections* [2], which only transfer information between streams that process *compatible* data. Both XFlow architectures outperformed their baselines (by up to 8.4%) when evaluated on the *AVletters*, *CUAVE* and *Digits* datasets, achieving state-of-the-art results.

## Model construction

The CNN × MLP architecture (shown on the right) takes as input a tuple `(x_img, x_mfcc)`: a 2D visual modality (the averaged video frames for a person saying a letter) and averaged 1D audio data corresponding to the same frames. The {CNN × MLP}–LSTM (shown below) processes the same kind of data, with the exception of each video frame/MFCCs pair being provided separately as input to the pre-concatenation streams. The crucial advantage of not having to average the data across more frames keeps the temporal structure intact and maintains a richer source of features from both modalities.

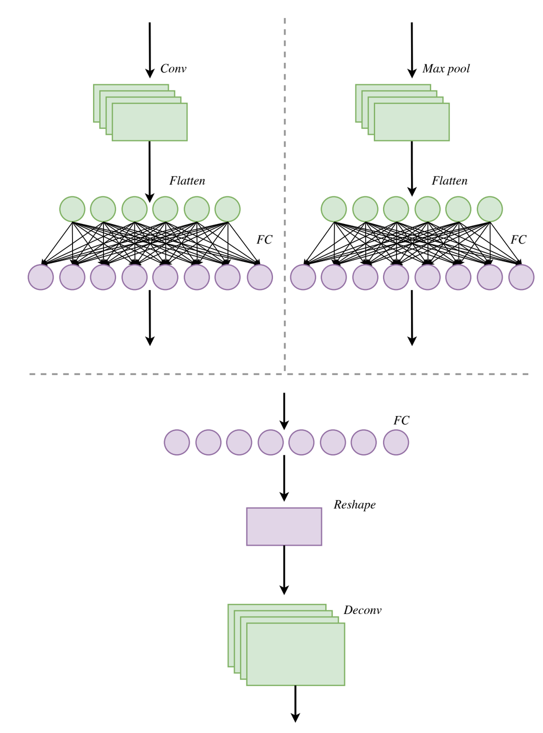



CNN × MLP model with cross- and residual connections. The feature extractor concatenates representations from both modalities.

## Cross-connections

The 1D⇝2D cross-connections take the output of a fully-connected layer and pass it through another layer of the same type, such that the number of features matches the dimensionality required for the *deconvolution* operation. We then apply the latter to the reshaped data and concatenate the result with the output of a {conv×2, max-pool} block. The 2D⇝1D cross-connections perform an inverse operation. Finally, residual connections are constructed in a similar manner.




*(Upper left:)* 2D⇝1D cross-connection. *(Upper right:)* 2D⇝1D residual. *(Bottom:)* 1D⇝2D cross-/residual.

## References

[1] C. Cangea, P. Veličković, and P. Liò. XFlow: 1D-2D Cross-modal Deep Neural Networks for Audiovisual Classification. *ArXiv e-prints*, September 2017.

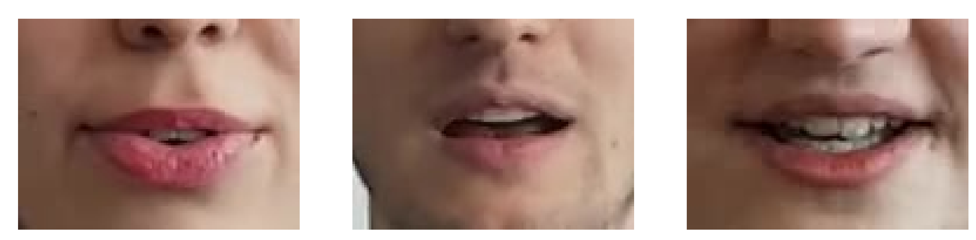
[2] P. Veličković, D. Wang, N. D. Lane, and P. Liò. X-CNN: Cross-modal Convolutional Neural Networks for Sparse Datasets. *ArXiv e-prints*, October 2016.

[3] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

[4] Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal, and Jiebo Luo. Deep multimodal representation learning from temporal data. *CoRR*, abs/1704.03152, 2017.

## Evaluation

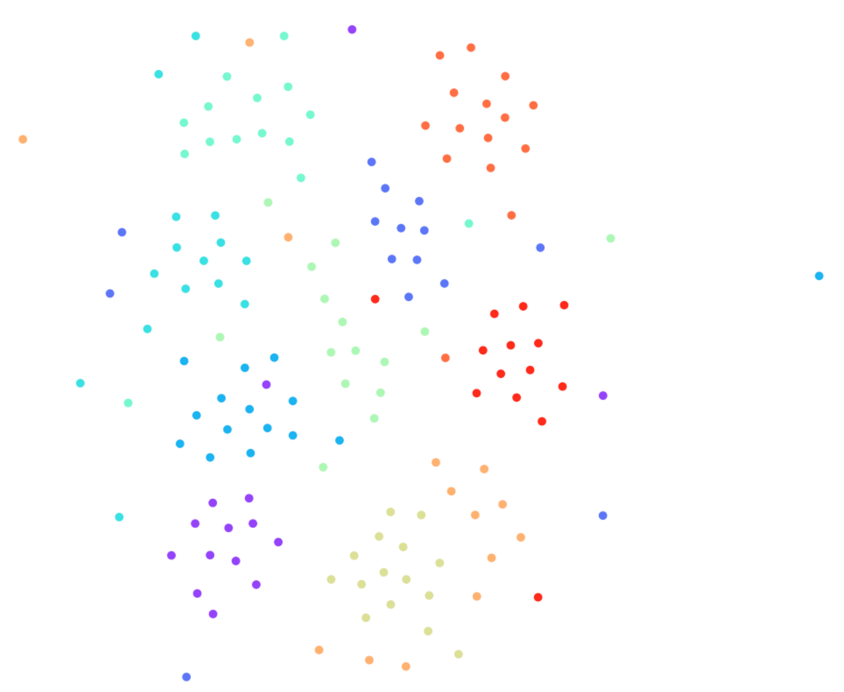
We evaluated the models using *AVletters*, *CUAVE* and the novel *Digits* datasets. *AVletters* contains 780 examples of 10 people saying each letter three times, distributed across 26 classes, whereas *CUAVE* has 36 people saying each digit five times. With 750 examples belonging to 10 classes (digits 0–9) and 15 people, *Digits* contains three different data types (video frames—a few examples are shown below, audio coefficients and spectrograms).



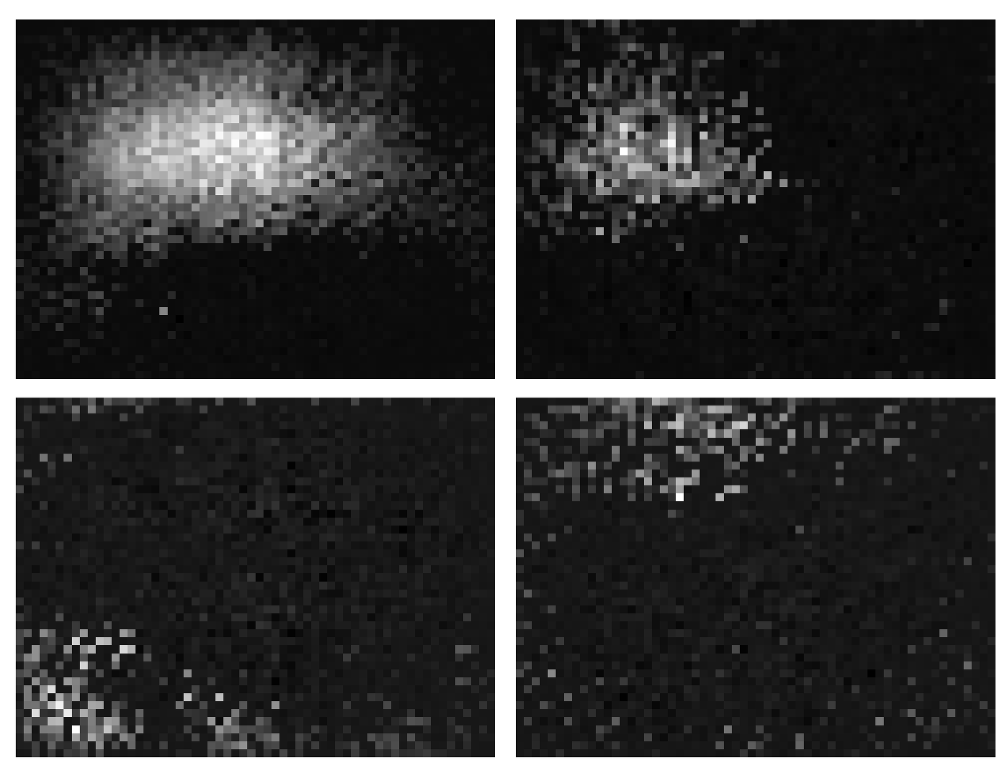
We used cross-validation and the holdout setup in [3] to compare the XFlow models to their baselines (without the cross-modal connections); folds correspond to disjoint groups of people. The {CNN × MLP}–LSTM outperformed CorrRNN, the state-of-the-art result in [4].

| | ***AVletters*** | | | | ***Digits*** | | | ***CUAVE*** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Baseline | XFlow | CorrRNN | *p*-value | Baseline | XFlow | p-value | Baseline | XFlow | CorrRNN | *p*-value |
| CNN × MLP | 73.1% | **74.0%** | — | 0.65 | 78.3% | **86.7%** | $\underline{2 \times 10^{-3}}$ | 90.3% | **93.5%** | — | $\underline{0.05}$ |
| {CNN × MLP}–LSTM (CV) | 78.1% | **85.6%** | — | $\underline{0.02}$ | 88.7% | **93.0%** | $\underline{1.2 \times 10^{-3}}$ | 96.9% | **98.8%** | — | $\underline{0.01}$ |
| {CNN × MLP}–LSTM (holdout) | 91.5% | **94.6%** | 83.4% | — | — | — | — | 96.1% | **96.9%** | 95.9% | — |

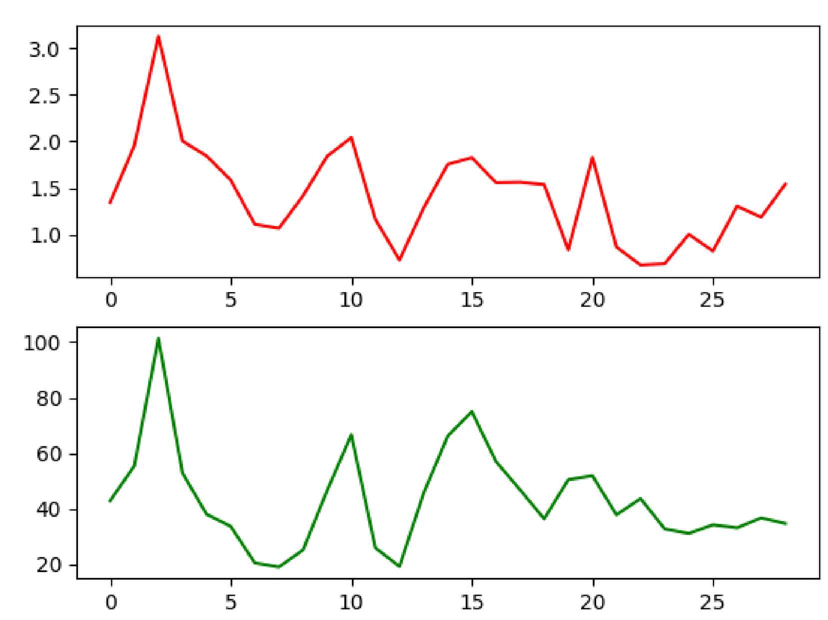
## Interpretability of cross-modal transformations (pre-trained on *Digits*)



*t*-SNE plot of the outputs of the second CNN × MLP 2D⇝1D cross-connection; the colours correspond to *Digits* classes.



Example outputs of the first {CNN × MLP}–LSTM residual connection.




Differences between the residual connection outputs (in red) and the MLP 1D inputs (in green).