1D $\leftrightsquigarrow$ 2D Cross-modality for deep audiovisual classification

Cătălina Cangea

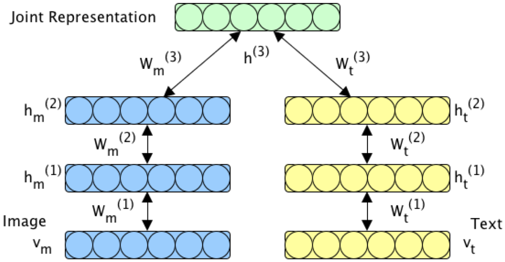
Computer Laboratory, University of Cambridge, UK

The problem

- Aim to improve classification performance of a multimodal recognition system
- Learn from multiple representations (images, speech, ...) of the same symbols (0–9, A–Z)

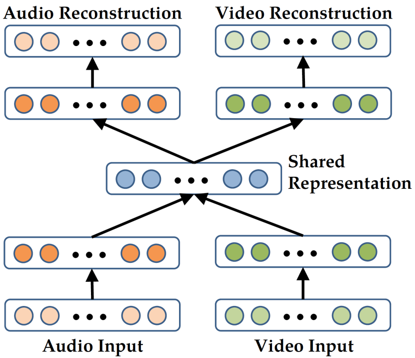
Previous approaches

Srivastava et al. (2012) — multimodal Deep Boltzmann Machine fusing images and text



Previous approaches

Ngiam et al. (2011) — bimodal deep autoencoders fusing audio and video



Cross-modality

- Only previously done after feature extraction
- ...but likely to increase classification performance if done *during* this step — exploit correlations
- Non-trivial between incompatible (both spatially and semantically) data types (audio/video)

Contributions

1. Three deep learning **architectures** with cross-modal feature extractors, each processing two modalities

2. A new high-quality audiovisual **dataset**

3. **Interpretability** of cross-modal exchanges $\rightarrow$ conclusions on mutual influence between feature extractors and data types

Models

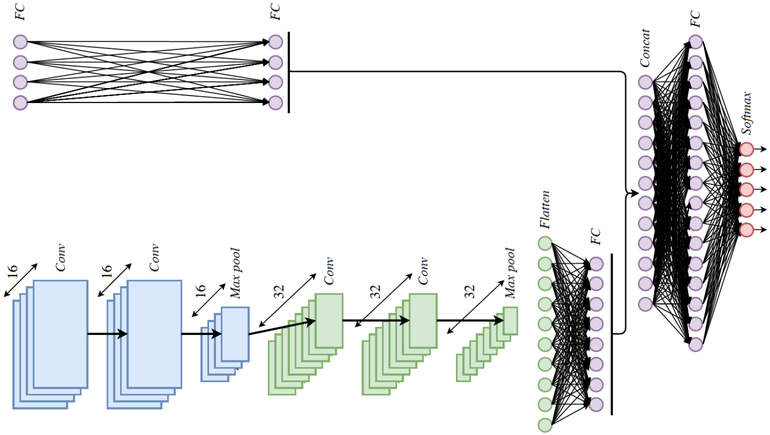
1. **CNN** $\times$ **MLP**: take as input video frames and MFCCs for the entire sequence;

2. **CNN** $\times$ **CNN**: video frames and spectrograms for the entire sequence;

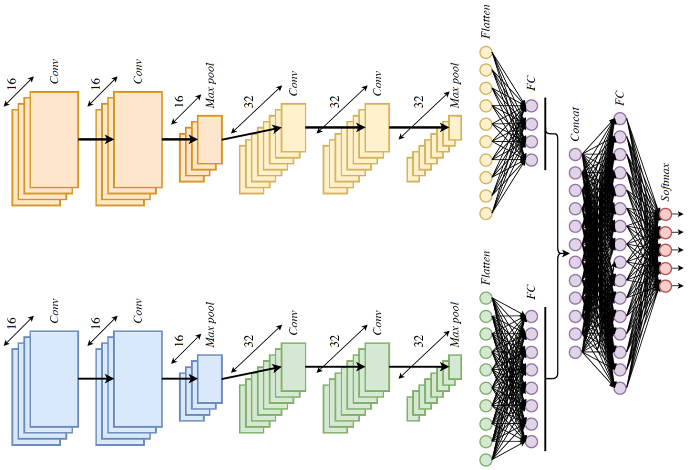
3. $\{$**CNN** $\times$ **MLP**$\}$–**LSTM**: video frames and corresponding MFCCs, frame by frame.

The first 2 models process fixed-length sequences; had to average examples across suitable windows, resulting in loss of information.

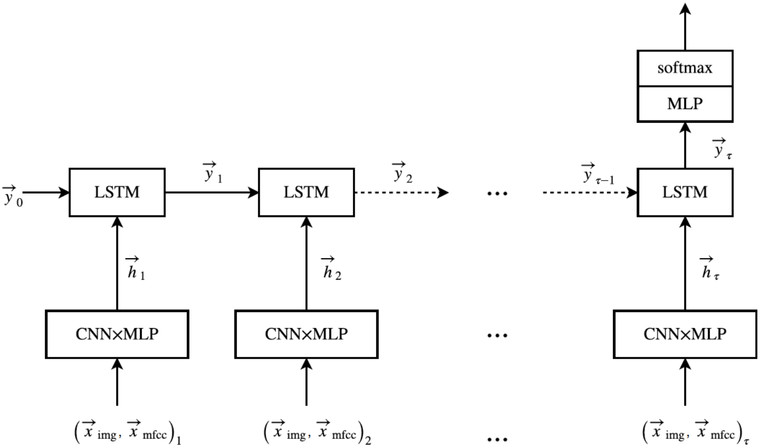
CNN × MLP baseline



CNN × CNN baseline

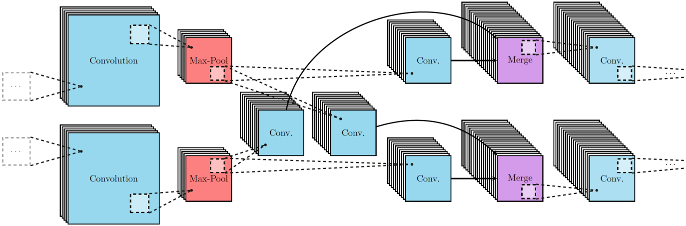


$\{$CNN $\times$ MLP$\}$–LSTM baseline



Cross-connections

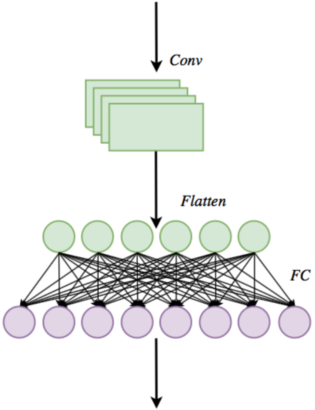
- Introduced by Veličković et al. (2016)
- Exchange feature maps between streams that process *compatible* data (e.g. YUV channels)

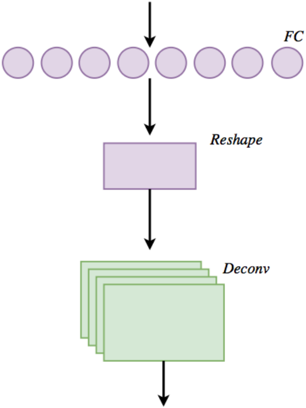


Non-trivial cross-connections

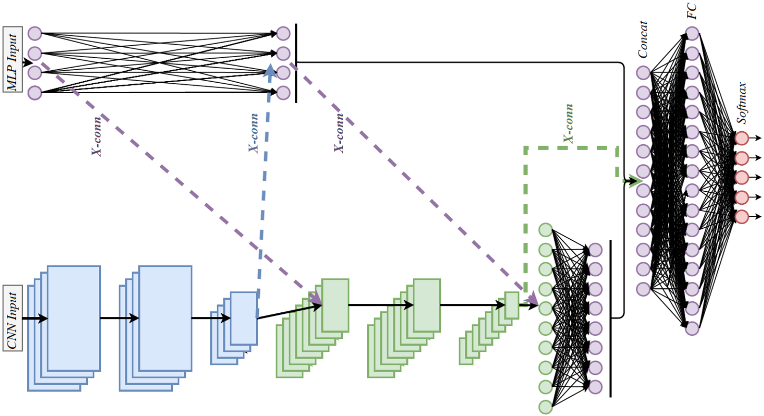
- 2D $\rightsquigarrow$ 1D: pass 2D features through a convolutional layer, flatten the result and send it to a fully-connected layer which produces 1D output
- 1D $\rightsquigarrow$ 2D: pass 1D features through a fully-connected layer, reshape the result and deconvolve it to obtain data in a matching shape for the other stream
- 2D $\rightsquigarrow$ 2D: carefully deconvolve to account for the differences in aspect ratio

Cross-connections for **CNN** × **MLP**



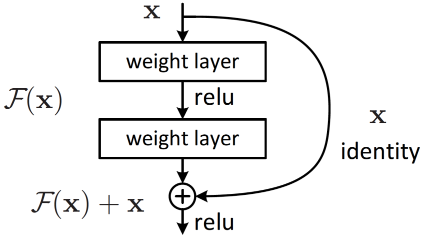


CNN × MLP with cross-connections



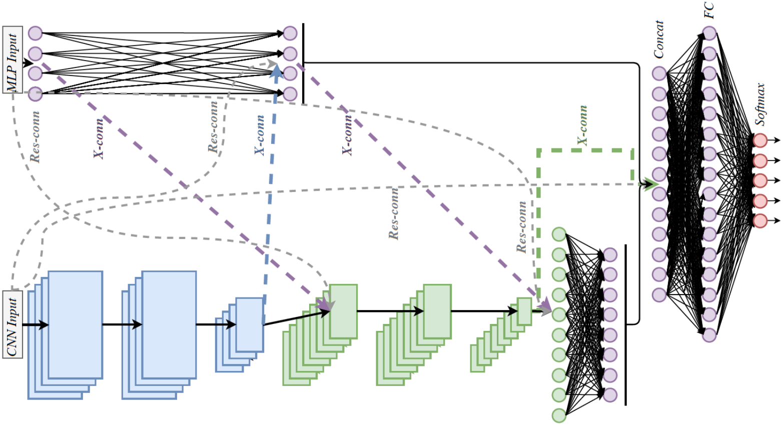
Residual connections

"Shortcut" connections introduced by He et al. (2016) to facilitate designing deep architectures



My work allows to shortcut inputs between incompatible streams in a straightforward manner.

CNN × **MLP** with cross-connections and residuals



Cross-connection regularisation

- Merging a stream with a cross-connection output increases the number of parameters in the next layer—need increased regularisation after the merging point (dropout from 0.25 to 0.5)
- *ReLU* activation used in all intermediate layers, but cross-connections use *PReLU* (parametric ReLU) to maintain information integrity:

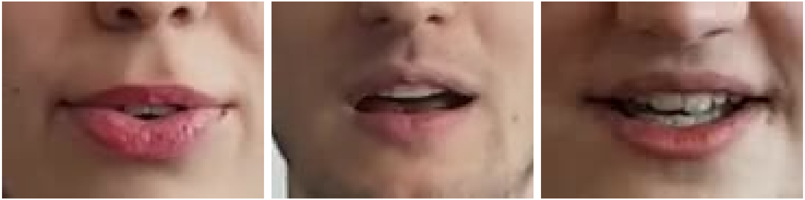
$$PReLU(x) = \begin{cases} \alpha x, & x \leq 0, \\ x, & x > 0, \end{cases}$$

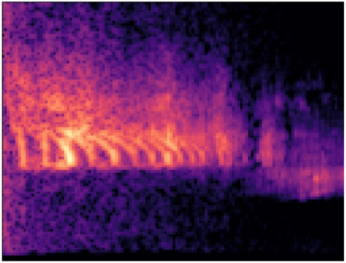
  where α is learnable (and always 0 for *ReLU*).

Digits dataset

- Existing datasets (AVletters, CUAVE) were either inaccessible or over-processed
- Collected data consisting of 750 high-quality examples of 15 people, each saying the digits 0–9 in 5 different tones
- Processed three modalities: video frames (2D), MFCCs (1D), spectrograms (2D)

Digits dataset





Results for AVletters

	Baseline	Cross-connected	p-value
CNN $\times$ MLP	73.1%	**74.0%**	0.65
$\{$CNN $\times$ MLP$\}$–LSTM	78.1%	**85.6%**	<u>0.02</u>

AVletters was over-processed, which resulted in a poor modality alignment exacerbated by window averaging—the only situation where the fixed-length model was not *significantly* better.

Results for CUAVE

	Baseline	Cross-connected	*p*-value
CNN $\times$ MLP	90.3%	**93.5%**	0.05
$\{$CNN $\times$ MLP$\}$–LSTM	96.9%	**98.8%**	0.01

Results for Digits

	Baseline	Cross-connected	p-value
CNN $\times$ MLP	78.3%	**86.7%**	$\underline{2 \times 10^{-3}}$
CNN $\times$ CNN	66.7%	**70.4%**	$\underline{5 \times 10^{-4}}$
{CNN $\times$ MLP}–LSTM	88.7%	**93.0%**	$\underline{1.2 \times 10^{-3}}$

Interpretability

- Adding cross-connections enables the modalities to interact more usefully towards building a stronger joint representation

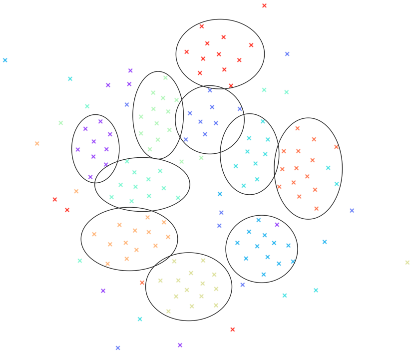
- Investigated the discriminative properties of cross-connections (2D $\rightsquigarrow$ 1D) and their ability to pass features between streams in a structurally interpretable manner (1D $\rightsquigarrow$ 2D)

t-SNE

- A dimensionality reduction method that preserves the notion of distance between the points in a high-dimensional feature space, allowing for detecting interpretable 2D clustering.

- Investigated outputs from a 2D $\rightsquigarrow$ 1D connection from the CNN $\times$ MLP model

t-SNE visualisation

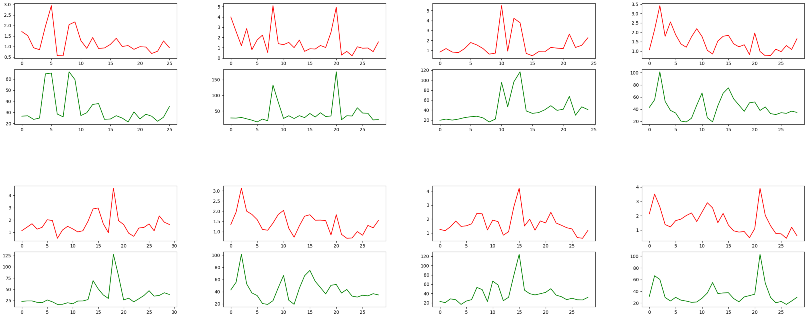


Visible clustering observed across the different classes (0–9).

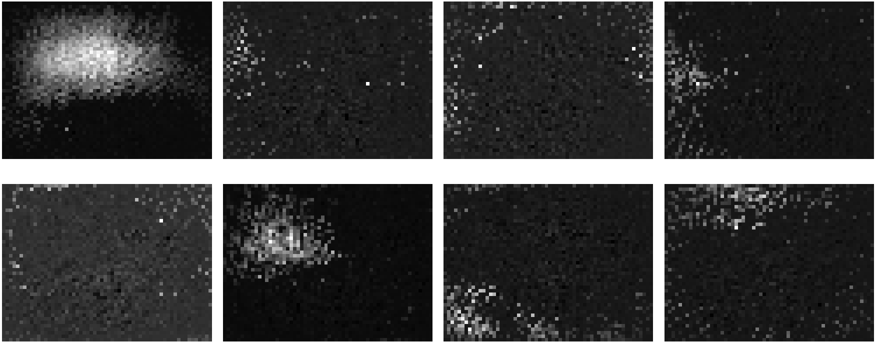
Structural interpretability

- Analysed a 1D $\rightsquigarrow$ 2D residual cross-connection from the $\{$CNN $\times$ MLP$\}$–LSTM model
- Plotted Euclidean distances (L^2 norms) between consecutive input sequences and the corresponding outputs of the residual connection
- Visualised activations of the cross-connection for several examples, across all timesteps

Euclidean distances



Activations



Conclusions

- Devised a novel way of exchanging information between fundamentally incompatible data types in the feature extraction stage, obtaining highly significant improvements in classification performance
- Created a new high-quality dataset that can be used for future multimodal research
- Made steps towards higher interpretability of multimodal learning
- *Work presented in a poster at the ARM Research Summit 2017 and during a presentation at the Workshop on Computational Models for Crossmodal Learning (CMCML), IEEE ICDL-EPIROB 2017.*

Thank you!

Questions?