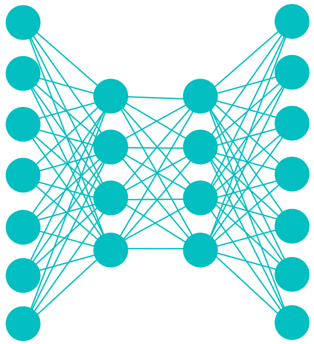
Lecture Notes for

Neural Networks and Machine Learning

Semi-supervised Loss

Next Time:
MML and MTL

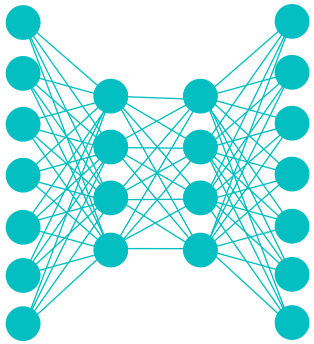
Reading: None



Lecture Notes for

Neural Networks and Machine Learning

Multi-task and Multi-Modal Learning



Logistics and Agenda

- Logistics
  - Lab one due today!
  - Special office hours today 1:30-3PM
- Agenda
  - Student Paper Presentation
  - Multi-modal and Multi-task
- Next Time
  - Multi-task demo and Town Hall
  - Finish Demos

Paper Presentation: Deep Fake Detection

Combining EfficientNet and Vision Transformers for Video Deepfake Detection

Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi

ISTI-CNR, via G. Moruzzi 1, 56124, Pisa, Italy
davidealessandro.coccomini@isti.cnr.it, nicola.messina@isti.cnr.it,
claudio.gennaro@isti.cnr.it, @fabrizio.falchi@isti.cnr.it

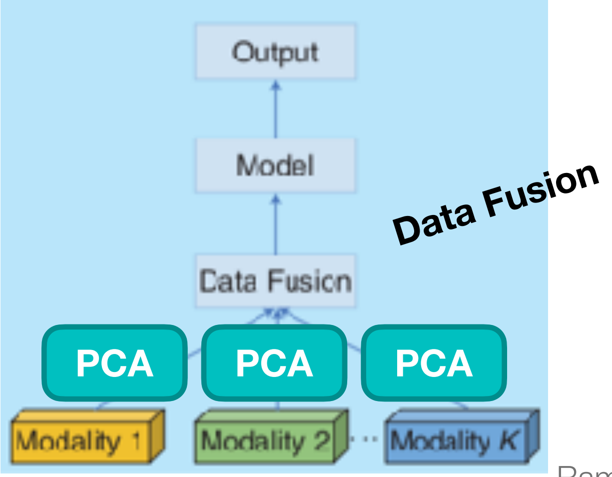
Multi-modal Review

Multi-modal == Multiple Data Sources

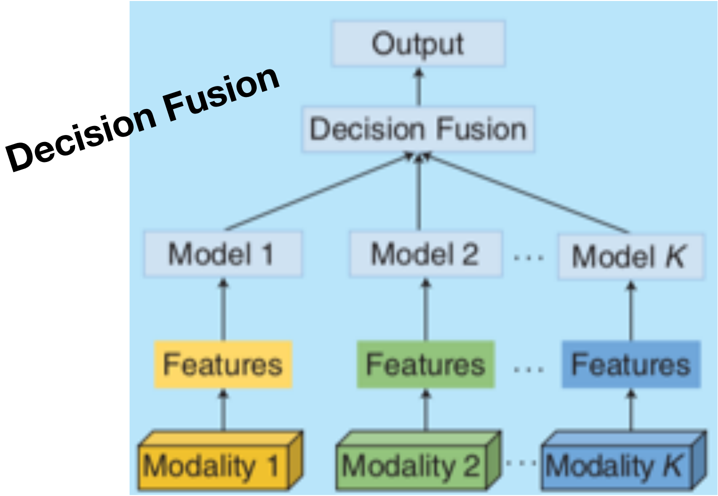
- **Modal** comes from the “sensor fusion” definition from Lahat, Adali, and Jutten (2015) for deep learning
- Using the Keras functional API, this is extremely easy to implement
  - … and we have used it since CS7324!
- But now let’s take a deeper dive and ask:
  - What are the different types of modalities that we might try?
  - Is there a more optimal way to merge information?
  - When? Early, Intermediate, and late fusion

Early and Late Stage Fusion

- **Early Fusion:** Merge sensor layers early in the process
- **Assumption**: there is some data redundancy, but modes are conditionally dependent
- **Problem**: architecture parameter explosion
  - Need dimensionality reduction



- **Late Fusion**: Merge sensor layers right before flattening
- Use Decision Fusion on outputs
- **Assumption**: little redundancy or conditional independence—just an ensemble architecture
- **Problem**: just separate classifiers, limited interplay

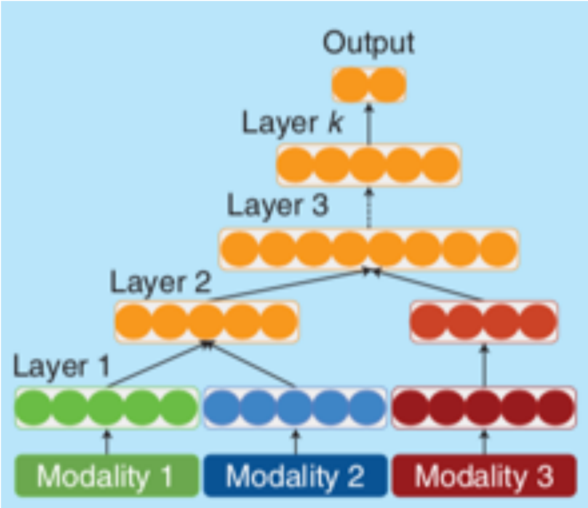


Ramamchandran and Taylor, 2017

Intermediate Fusion

- Merge sensor layers in soft way
- **Assumption**: some features interplay and others do not
- **Problem**: how to optimally tie layers together?

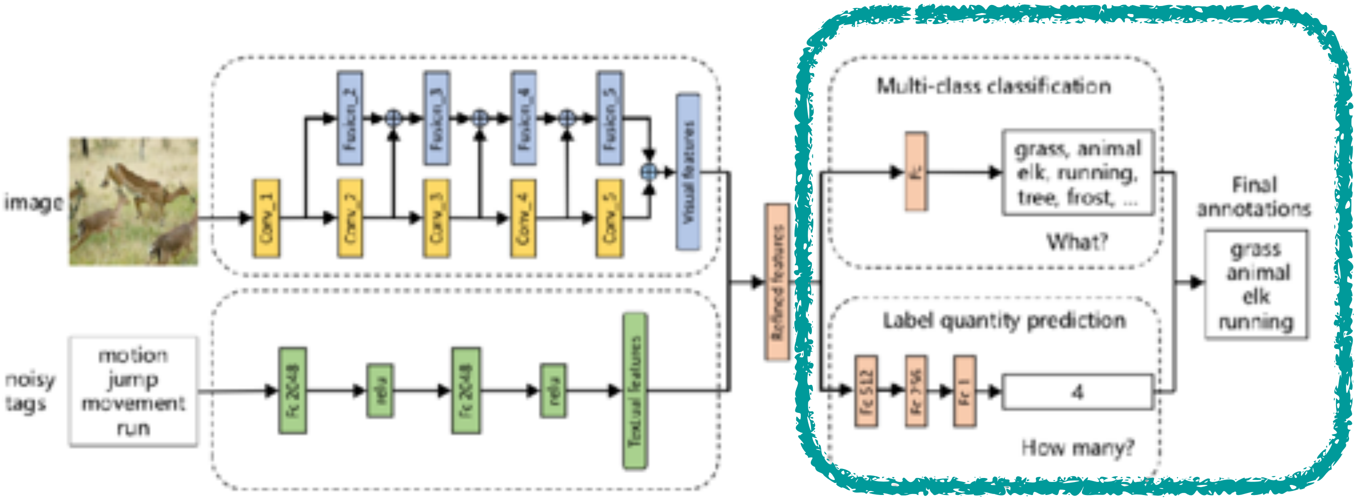
1. Stacked Auto-Encoders
   [Ding and Tao, 2015]
2. Early fuse layers that are correlated
   [Neverova *et al.* 2016]
3. Fully train each modality merge based on criterion of similarity in activations
   [Lu and Xu 2018]
4. Granger Cluster data in each modality and combine [Sylvester *et al.* 2023]



Ramamchandran and Taylor, 2017

Multi-modal Merging

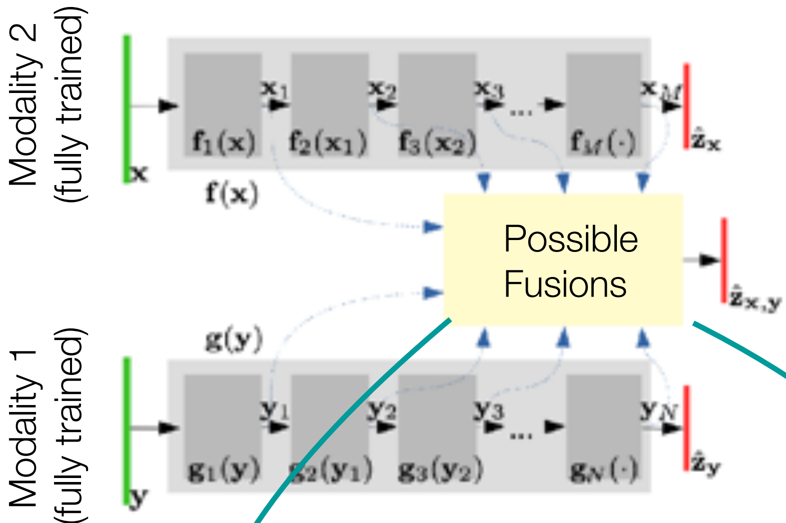
- **Still an open research problem**
- How to develop merging techniques that
  - Can handle exponentially many pairs of modalities
  - Automatically merge meaningful modes
  - Discard poor pairings
  - Selectively merge early or late (or dynamically)



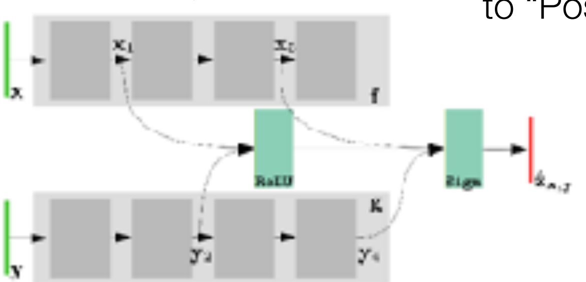
https://arxiv.org/pdf/1709.01220.pdf

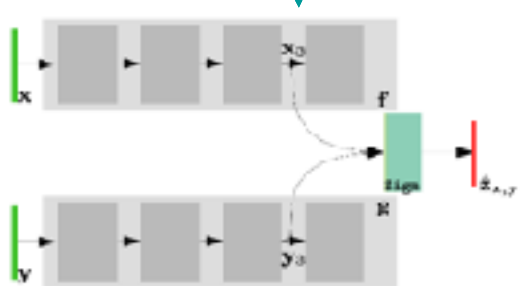
Most current methods are still ad-hoc

Neural Architecture Search for Mode Fusion



1. Sample blocks at Random
2. Train by fine tuning
3. Evaluate fitness
4. Add trained blocks back
   to “Possible Fusions”

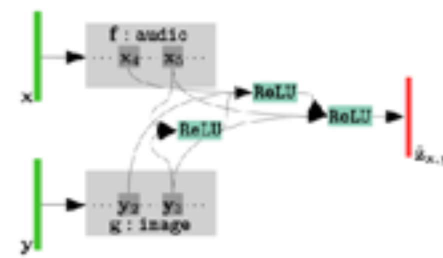




Genetic Algorithm

1. Sample new candidates
2. Evaluate fitnesses
3. Mutate and Crossover
4. Keep the best solutions
5. Repeat

Very computational when starting, because candidates are all untrained.
However, as more blocks start from “mostly trained” positions, training time reduces.



Found solution for AV-MNIST
x4, x5 and y2, y3

Pérez-Rúa, Juan-Manuel, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. "Mfas: Multimodal fusion architecture search." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6966-6975. 2019.

		Assistment 09-10			Oli Statics 2012		
		r2	AUC	wAUC	r2	AUC	wAUC
SM	DKT	0.1628	0.7326	0.7367	0.4106	0.8819	0.8622
	DKVMN	0.1507	0.7299	0.7354	0.3557	0.8793	0.8614
	SAKT	0.1541	0.7285	0.7223	0.3116	0.8373	0.8261
	NAS cell	0.1678	0.7364	0.7408	0.4169	0.8844	0.8661
MM	DKT + SC	0.1743	0.7371	0.7441	0.4316	0.8884	0.8734
	DKT + FS	**0.1844**	0.7454	0.7493	0.4239	0.8863	0.8651
	MFNAS	0.1829	**0.7458**	**0.7545**	**0.4348**	**0.8902**	**0.8779**

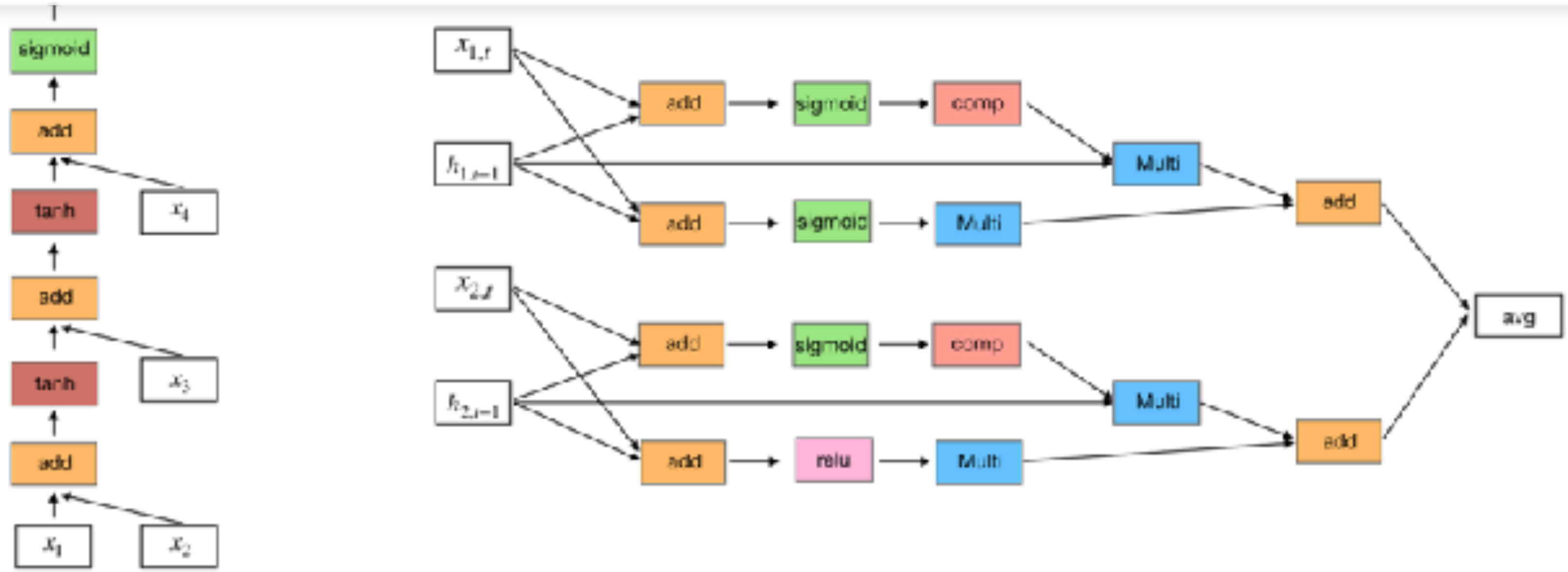


Fig. 5 The discovered best architectures for knowledge tracing. **Left:** Multimodal fusion search (FS) and use the fixed LSTM recurrent cell (Fig. 3). **Right:** Extend the sub-graph sampling to multimodality (Fig. 4). Here *add* stands for element wise addition. *tanh* is the hyperbolic tangent function, *sigmoid* is the logistic function, *relu* is the rectified linear activation function, *Multi* is the dot production and *comp* indicates (1 - c) operation

Approaches with Deep Learning

- Latent Space Transfer (universality)
  - From another domain, map to a similar latent space for the same task
  - Useful for unifying data based upon a new input mode when old mode is well understood
    - for example, biometric data
    - **I have never seen a research paper on this…**