Sensory Integration in Deep Neural Networks

Marek Dobeš*, Rudolf Andoga, Ladislav Főző****

* CSPV SAV, Karpatská 5, 04001 Košice, Slovak Republic, dobes@saske.sk

** Faculty of Aeronautics, Technical University of Košice, Rampová 7, 041 21 Košice, Slovak Republic, rudolf.andoga@tuke.sk; ladislav.fozo@tuke.sk

Abstract: Two unimodal deep networks and one multimodal deep network are created to test for possible mechanisms of sensory integration that may shed more light on how sensory integration is carried out in biological organisms. One unimodal network is provided with pictures and the other with mel-spectrograms created from sounds. Adapted pre-trained VGG16 network was used for unimodal networks. After training consisting of 30 epochs and repeated for 100 runs the unimodal networks achieved an average accuracy of 0.57 and 0.73 respectively. The multimodal network received processed features from both unimodal networks and after training consisting of 30 epochs and repeated for 100 runs outperformed both unimodal networks with the average accuracy of 0.79. Next, noise was applied to the test data to see how unimodal and multimodal networks compare in noisy environments. Unimodal networks achieved an average accuracy of 0.63 and 0.69 respectively. Again, the multimodal network outperformed both unimodal networks with an average accuracy of 0.73. Pre-trained networks were used and limited training data were provided to the networks to simulate conditions similar to animal brains.

Keywords: sensory integration; deep learning; neural network

1 Introduction

Biological organisms evolved to make use of different modalities – types of stimuli that they react to – such as visual, audio, chemical and others. Sensory integration in living organisms is an important tool that enables animals to better differentiate between objects and get information in noisy environments [1]. Though many studies have been done on neurophysiology of multimodal integration, mostly in invertebrates such as Drosophila [2] or wolf spider [3], we still do not know precisely how such integration is done on a neuronal level, not even in simple organisms like C. Elegans [4]. Computational modelling may provide insights into these processes as it allows for experiments that are not possible with living brains either because they are not feasible or ethically viable [5].

Deep neural networks thank, in part, their popularity because they are able to identify patterns in a way similar to how human brains do. They are successful in recognising pictures, sounds, and other data [6]. However, most applications are in the domain of one modality (visual, audio, or other). When two or more unimodal networks are to be combined into a multimodal network to improve classification results there are more approaches that can be used. Probably the simplest approach is to do a weighted average of network results [7]. Most deep neural networks used for classification have a top softmax layer that represents probabilities that the object provided to the network belongs to a certain class. Probabilities from unimodal networks are then weighted and averaged to provide final probabilities. While this approach can improve classification it does not use all the advantages that multimodal integration has over unimodal classification. The power of multimodal integration using deep neural networks lies in the fact that data that were not used in one modality can be useful when combined over more modalities. Features that have not been used in simple classification tasks can be utilised in connection with features from other networks [7]. A more effective approach seems to lay in building a multimodal deep network that takes as its inputs features produced by unimodal networks [7]. This network is then trained on the multimodal data and should classify the objects more precisely. Typically a dense layer is used for integrating the inputs [8], although other approaches are tested such as using lateral connections and self-organisation [9], fully convolutional neural networks [10], or sequential late fusion [11].

In this study, we aim to explore how two unimodal networks can be fused into a multimodal network using deep neural networks and compare the performance of unimodal and multimodal networks. We would like to bring new evidence to how these networks perform in comparison with one another and how do they respond to noisy stimuli. In this study, in contrast with other studies mentioned above, we use pre-trained deep neural networks. Such a setup has two similarities to the neural systems of living organisms. First, living organisms have pre-wired neural systems that evolved during their phylogenesis. Second, pre-wired (or pre-trained) networks are able to learn even from a few examples as in natural settings it would be very disadvantageous to be able to learn only from thousands of examples.

2 Methodology

2.1 Architecture

Keras toolbox (keras.io) running under TensorFlow (tensorflow.org) in Python environment has been used for simulation experiments.

As was mentioned in the introduction, the aim is to inspect how a multimodal network working with two sets of features from unimodal networks perform in

comparison to unimodal networks. The model consists of two unimodal deep neural networks and one multimodal deep neural network. The first unimodal network is designed for the classification of visual data and the second unimodal network is designed for the classification of audio data converted to mel-spectrograms. To allow for faster training and lower the need for large samples we use Imagenet network [12] embedded into Keras environment as VGG16 for both unimodal networks. The embedding allows us to use the network directly and modify it easily for our purposes. The top classification layer was removed and substituted with a set of dense, dropout and dense layer with softmax activation to allow it to adapt to our data. The layers were frozen for training except for the top three layers to allow transfer learning on our dataset. As inputs we use 150*150*3 RGB images and each network outputs a vector of length 8192 that represents high-level features of the input image.

We concatenate the output vectors into a 16384 vector to represent both modalities and feed this vector into a multimodal network. The architecture of the multimodal network consists of an input layer, a dense layer with 1024 fully connected neurons, a drop-out layer, and a softmax classification layer. A block diagram is shown in Figure 1.

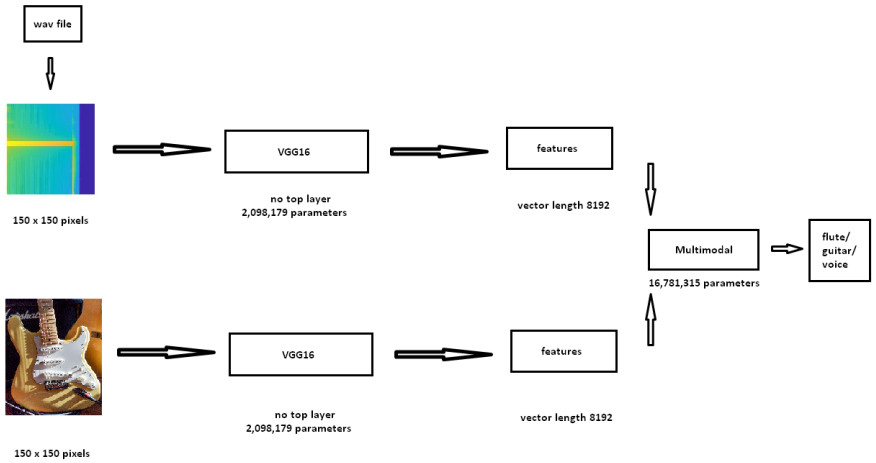

Figure 1
Block diagram of used architecture

2.2 Data

We use two modalities of data and three categories that are the same in each modality. We use data from two public databases. For visual data we use photographs from UnSplash (unsplash.com) and for audio data, we use audio samples from NSynth [13] that are converted into mel-spectrograms. We use three categories – voice – with a sound sample of singing and a photograph of a singing

person/s; guitar – with a sound sample of a guitar sound and a photograph of a guitar and flute – with a sound sample of a flute and a photograph of flute/s. The distribution of data is shown in Table 1.

Table 1
Distribution of data

	Flute	Flute	Guitar	Guitar	Voice	Voice
	Sound	Picture	Sound	Picture	Sound	Picture
training	6	6	14	14	11	11
test	6	6	14	14	11	11
validation	3	3	3	3	3	3

Sample picture and sample mel-spectrogram are found in Figure 2a and 2b. 31 samples of every modality are used for training and 31 samples for testing. For an experiment with noisy data, we use the same samples with added Gaussian noise (zero-mean white noise with variance 0.01). For an example of a noisy picture and noisy mel-spectrogram, see Figure 2c and 2d.

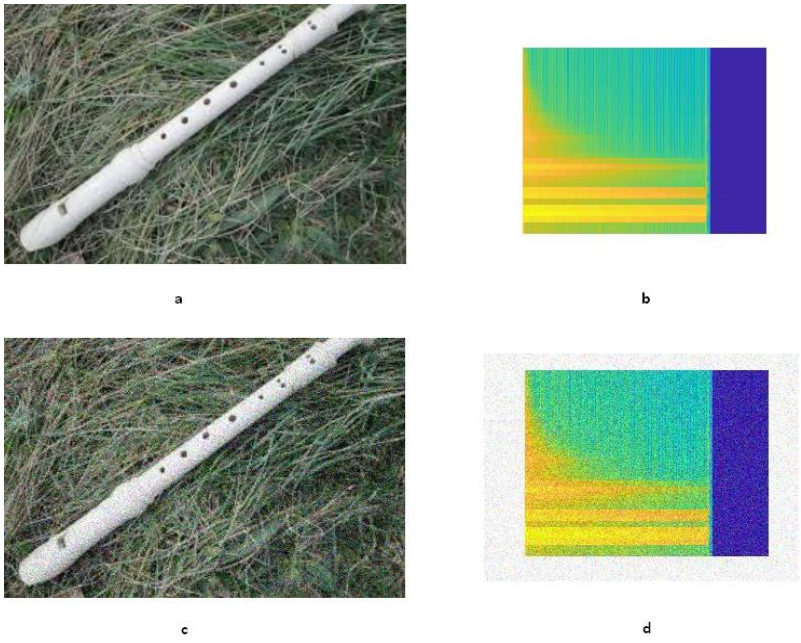

Figure 2
Sample data. a) picture, b) mel-spectrogram, c) noisy picture, d) noisy mel-spectrogram

2.3 Training

Unimodal and multimodal networks are trained using RMSprop optimizer, for loss we use sparse categorical crossentropy, the number of epochs is 30. Figure 3 shows the development of averaged training accuracy (acc; https://github.com/keras-team/keras/blob/68dc181a5e34d1f20edabe531176b3bfb50001f9/keras/engine/training.py#L375) and training loss (sparse categorical crossentropy; keras.io/api/losses/) across 100 runs.

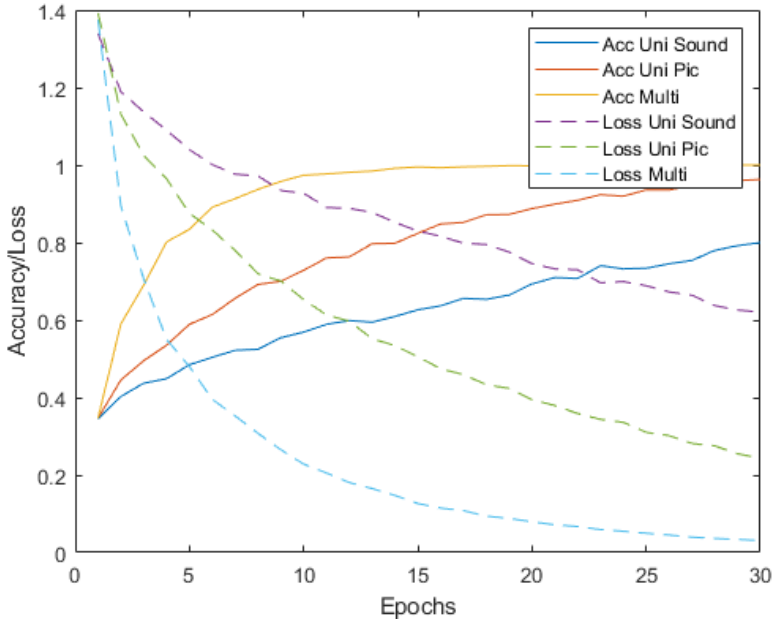

Figure 3
Average training accuracy and training loss

Training graph shows that multimodal network achieves better accuracy and lower loss over time and outperforms both unimodal networks.

3 Results

First, we tested both unimodal and multimodal networks to see whether multimodal network outperforms the unimodal networks. As deep neural networks use stochastic processes we repeated training and measured accuracy for 100 runs. We found the average accuracy for the first unimodal network (audio) to be 0.57, for the second unimodal network (visual) 0.73, and for the multimodal network

0.79. Graph of individual runs is shown in Figure 5. Multimodal network performed better than any single unimodal network.

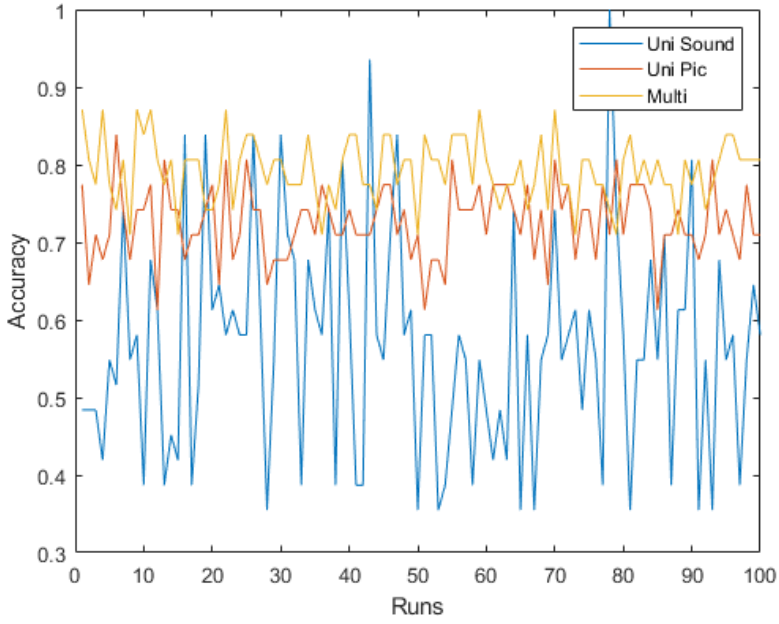

Figure 4

Accuracy of unimodal and multimodal networks over 100 runs. First unimodal network - blue, second unimodal network - red, multimodal network - orange

Next, we were curious about how the network behaved under noisy inputs. As we mentioned in the Introduction, multimodal networks should enhance recognition when inputs are noisy.

We used the same networks, parameters, and training data as in the previous experiment. Before testing the networks we added Gaussian noise to the inputs. We used zero-mean white noise with a variance 0.01. Again, we repeated training and testing for 100 runs and measured the accuracy of the networks. We found the average accuracy for the first unimodal network (audio) to be 0.63, for the second unimodal network (visual) 0.69, and for the multimodal network 0.73. Graph of individual runs is shown in Figure 5. Again, the multimodal network outperformed both unimodal networks although its performance was not as good as on the data without noise. The accuracy for the second unimodal network (visual) also decreased. Unexpected was a slightly better performance of the first unimodal network (audio) for noisy inputs. This may be due to stochastic fluctuation or to some, yet unknown, aspect of the interaction of deep neural networks with spectrogram data.

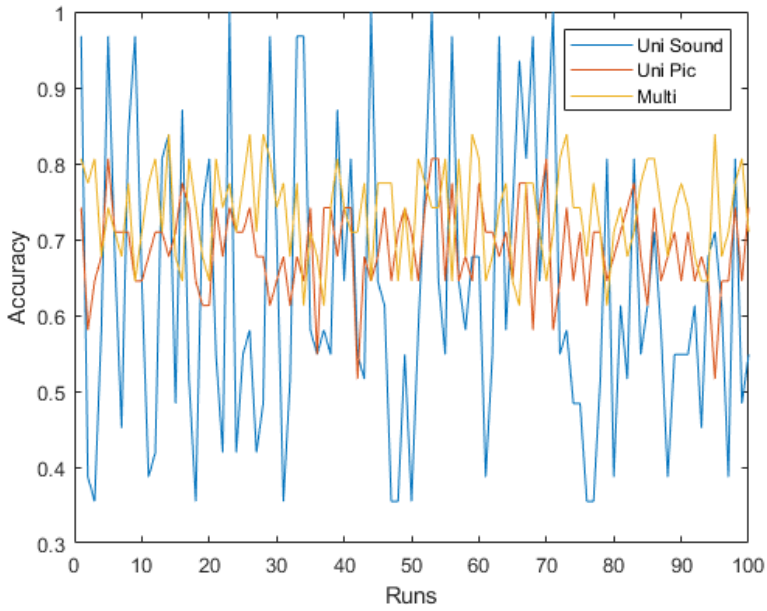

Figure 5
Accuracy of unimodal and multimodal networks with noisy test data over 100 runs. First unimodal network - blue, second unimodal network - red, multimodal network - orange

We created the confusion matrices for both unimodal and multimodal networks and for both not-noisy and noisy test data. In the tables, we summed the classification results for 100 runs. The results are shown in Table 2.

Table 2
Confusion matrices

Unimodal network - sound; not-noisy			
Input/Prediction	Flute	Guitar	Voice
Flute	32	16	12
Guitar	0	19	121
Voice	0	0	110
Unimodal network - picture; not-noisy			
Input/Prediction	Flute	Guitar	Voice
Flute	26	32	2
Guitar	4	102	34
Voice	0	13	97
Multimodal network; not-noisy			
Input/Prediction	Flute	Guitar	Voice
Flute	47	13	0
Guitar	7	93	40
Voice	0	0	110

Unimodal network - sound; noisy			
Input/Prediction	Flute	Guitar	Voice
Flute	42	8	10
Guitar	0	62	78
Voice	0	2	108
Unimodal network - picture; noisy			
Input/Prediction	Flute	Guitar	Voice
Flute	26	30	4
Guitar	2	91	47
Voice	0	14	96
Multimodal network; noisy			
Input/Prediction	Flute	Guitar	Voice
Flute	46	14	0
Guitar	2	72	66
Voice	0	1	109

Confusion matrices confirm our previous results. A multimodal network shows better classification results than any of unimodal networks using both not noisy and noisy test data. Upon closer inspection, we see that the worse performance of the unimodal network for sounds was primarily caused by misclassification of class Guitar in a not-noisy situation. We may only speculate that the classification of mel-spectrograms is a more complex task and adding the noise somehow tipped the algorithm towards better performance. This issue remains open for further research. On the other hand, this result underscores the fact that the advantage of multimodal networks is the fact that even if some of the unimodal networks do not work perfectly, their drawbacks are compensated using data from other unimodal networks when integrated into a multimodal network.

Conclusions

Sensory integration is an advantage for living organisms as it enables them to better classify objects and extract data from noisy environments. Deep neural networks are in many respects similar to biological ones and thus can help us to obtain insights using experiments that would otherwise be not feasible. In this study, we wanted to test whether a multimodal network integrating inputs from two unimodal networks representing two modalities can outperform these networks. Besides, we wanted to test whether such performance is possible using only a limited number of stimuli as is the norm in living systems. We also feel that it is important that our study has been done using open-platform software and data and we described the details of architecture and training parameters so that this study can be used for inspiration and subsequent research for other scientists.

Our results show, that the multimodal network outperformed both unimodal networks. Furthermore, it outperformed them also when tested on noisy data.

Our simulations show that the superior performance of the multimodal network does not have to hold for every single run. One may argue that this would make the multimodal network unusable in natural settings. We, however, think that it may well be possible that biological networks may also work this way – the robustness of the network does not lay in how it performs every single run, but the research indicates that brains operate in a statistical fashion [14]. When an animal sees an image or hears a sound, the activity of visual/audio neurons does not stop immediately, but reverberates and thus enables the animal to make use of the statistical properties of neural signal [15].

We feel that such interconnections between neuroscience and computational modeling may be fruitful for both research fields and may bring further insights into how animal (and human) brains work.

Acknowledgement

This work was supported by VEGA 2/0043/17 grant.

References

[1] CORCORAN, Aaron J.; MOSS, Cynthia F. Sensing in a noisy world: lessons from auditory specialists, echolocating bats. Journal of Experimental Biology, 2017, 220.24: 4554-4566

[2] HU, Chun, et al. Sensory integration and neuromodulatory feedback facilitate Drosophila mechanonociceptive behavior. Nature neuroscience, 2017, 20.8: 1085-1095

[3] KOZAK, Elizabeth C.; UETZ, George W. Cross-modal integration of multimodal courtship signals in a wolf spider. Animal cognition, 2016, 19.6: 1173-1181

[4] GHOSH, D. Dipon, et al. Multisensory integration in C. elegans. Current opinion in neurobiology, 2017, 43: 110-118

[5] URSINO, Mauro; CUPPINI, Cristiano; MAGOSSO, Elisa. Sensory fusion: A neurocomputational approach. In: 2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI). IEEE, 2016, pp. 1-6

[6] BENGIO, Yoshua; GOODFELLOW, Ian; COURVILLE, Aaron. Deep learning. Massachusetts, USA:: MIT press, 2017

[7] CHOLLET, F. Deep learning with python, Vol. 1, Greenwich, CT: Manning Publications CO, 2017

[8] AL-DULAIMI, Ali, et al. A multimodal and hybrid deep neural network model for remaining useful life estimation. Computers in Industry, 2019, 108: 186-196

[9] JAYARATNE, Madhura, et al. Bio-inspired multisensory fusion for autonomous robots. In: IECON 2018, 44th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2018, pp. 3090-3095

[10] BHANDARI, Abinash, et al. Object detection and recognition: using deep learning to assist the visually impaired. Disability and Rehabilitation: Assistive Technology, 2019, 1-9

[11] MODI, Anitha; SHARMA, Priyanka. SeLF: A Deep Neural Network Based Multimodal Sequential Late Fusion Approach for Human Emotion Recognition. In: International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2019, pp. 275-283

[12] DENG, Jia, et al. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009, pp. 248-255

[13] ENGEL, Jesse, et al. Neural audio synthesis of musical notes with wavenet autoencoders. In: International Conference on Machine Learning. PMLR, 2017, pp. 1068-1077

[14] DAYAN, Peter; ABBOTT, Laurence F. Theoretical neuroscience: computational and mathematical modeling of neural systems. Computational Neuroscience Series, 2001

[15] DEVORE, Sasha; DELGUTTE, Bertrand. Effects of reverberation on the directional sensitivity of auditory neurons across the tonotopic axis: influences of interaural time and level differences. Journal of Neuroscience, 2010, 30.23: 7826-7837