Towards Life Long Learning: Multimodal Learning of MNIST Handwritten Digits

Eli Sheppard*†, Hagen Lehmann†, G. Rajendran‡, Peter E. McKenna‡, Oliver Lemon*†, Katrin S. Lohan*†

*Edinburgh Centre for Robotics,
†Department of Computer Science, School of Mathematical and Computer Science, Heriot-Watt University, Edinburgh,
‡Department of Psychology, School of Social Sciences, Heriot-Watt University, Edinburgh

***Abstract*—Robots need to be able to continually learn from natural interactions, which are inherently multimodal or multi-sensory (e.g. hearing, vision). Here, we report the development of a deep multimodal autoencoder, capable of unsupervised learning. We created a joint vector space combining labels and images, using a multimodal autoencoder, which maps from images to labels and vice versa. We achieved a label prediction accuracy of 96.9% in the image-only testing condition, where the state-of-the-art is 99.79% and uses a committee of 5 neural networks. In robotics sensory data, is often cheap whilst computation is expensive, thus multimodal systems which require less computation to achieve comparable recognition rates are highly desirable.**

***Index Terms*—Deep Learning, Neural Networks, Multimodal Learning, Developmental Learning**

I. Introduction

Recent work in the deep learning community has focused on unimodal symbol processing, for example object recognition [1]. Whilst this approach has achieved excellent results for recognising objects, it does not provide a method for linking back to the sensory percepts which produce the recognised object class. Further to this, these methods are highly susceptible to noise and easily fooled [2].

From the psychology literature, we can see strong support for multimodal symbol processing [3]. Barsalou argues that cognition cannot occur through the processing of unimodal symbols alone, but occurs through a combination of multimodal symbol processing and simulation.

In this work we take a bio-inspired approach as it has been shown by other work that this may provide a more effective learning strategy [4] [5] [6].

With the paradigmatic shift in the definition of the nature of artificial intelligence from the traditional view of symbolic reasoning towards the idea of embodied cognition [7], [8], the central idea became that cognition emerges as the result of sensory-motor feedback loops when an agent interacts with its environment. In humans these feedback loops are necessarily multimodal due to the different sensors that we use to explore our environment e.g. tactile, visual, auditory, etc.

One characteristic of multimodality is the interrelating of multiple simultaneous representations across sensory inputs, in which the different sensory experiences are time-locked and correlated [10]. As a consequence the different sensory systems can inform one another without explicit external instructions [11]. Infants can commonly be observed watching their hands in front of their faces while manipulating objects [12]. In this way the tactile, visual and proprioceptional experiences with these objects are correlated.

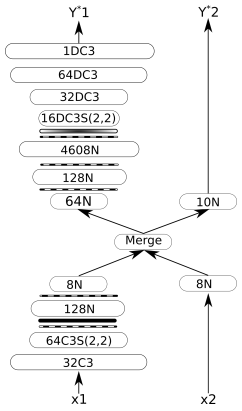

Fig. 1. Bimodal Autoencoder. Layers are labeled using the notation style of [9]. Layers marked C, DC,and N are convolution, deconvolution, and fully connected respectively. Striped bars represent dropout layers, solid bars represent flatten and the graded bar represents reshaping. S refers to (fractionally)strided convolutions which either downscale or upscale the image in convolutional and deconvolutional layers respectively.

Multimodality within an autoencoder therefore allows life long learning by removing the need for labeling of inputs and instead training the system to correlate data from different modalities. In the simple example presented here, this may be less apparent, however if the system were extended to combine sound and images, one could see how the time-locked sounds can be used as labels for the images and vice-versa.

Here we focus on the problem of label and object learning from multimodal sensory input. In this first work we will follow the approach presented by [4] and [13] using two modalities, supporting the reproducibility of the input features, this will be one of our measures representing the quality of what has been learnt by our system. We present a deep neural network architecture capable of learning a multimodal representation of MNIST handwritten digits [14].

Our network learns to anchor both images and labels to a continuous latent representation. Anchoring is the process of linking a symbol to a sensory percept [15] [16] [17] [6], in our case, the latent representation of a percept is its symbolic representation.

Since direct exploration of the quality of the internal representation learnt by the model is difficult, we use classification accuracy and reconstruction quality as secondary measures, to determine if a viable multimodal representation is achievable with this architecture i.e. if we can learn the same (or similar) representations for percepts of the same thing from different modalities.

We compare two different methods for merging multimodal data within the context of a deep multimodal autoencoder architecture.

II. Background

The area of multimodal learning has seen a large variety of research, looking at the combination of different modalities and the effects of these combinations on recognition accuracy, input denoising and reconstruction of missing modalities. Research has been carried out on both humans and robots to explore these effects.

A. Multimodal Speech Recognition

A.G. Samuel [18] found that humans reconstruct missing phonemes when hearing real words or pseudowords (fake words that are highly similar to real words) but not when hearing entirely fake words. This suggests that humans utilise lexical knowledge to reconstruct the missing phonemes. This supports the notion of multimodal symbol processing as being central to human cognition; if unimodal symbol processing were occurring, non-words would show the same level of reconstruction as real words, as reconstruction would occur in the acoustic modality, not the lexical.

Until the 1980s, research in traditional theories of cognition viewed knowledge as residing in a semantic memory system separate from the brains modal systems for perception (e.g., vision, audition), action (e.g., movement, proprioception), and introspection (e.g., mental states, affect) [3]. Cognitive science typically assumed a view of the mind as an abstract information processing system, in which our sensory and motor systems served a peripheral role conveying information to and from a central cognitive processor (the brain) where high level abstract thinking took place. This is in direct opposition to the findings of [18].

Ngiam et al. [4] developed a bimodal speech recognition system using Restricted Boltzman Machines (RBM). It successfully learnt to classify spoken digits using a combination of video and audio data. Whilst the classification accuracy with clean audio was lower using a bimodal RBM than using an audio only RBM, the combination of audio and visual information outperformed audio only when the audio was noisy.

Further to this, Ngiam et al. were able to show that their model could replicate the McGurk effect [19], an audio-visual illusion where the sound /ba/ combined with a visual /ga/ is heard by most humans as /da/. This lends credence to Barsalou's [3] belief that humans perform multimodal symbol manipulation. Therefore, if we want robots to be able to think and learn like humans, they too must process multimodal symbols.

B. Image and Text Processing

Silberer et al. [13] demonstrated the ability of deep bimodal autoencoders to reproduce textual and visual attributes. However, they did not go as far as reproducing actual images and text. This work presents a system which can reproduce images and categorical labels, thus building on their work by expanding their architecture to a new domain.

III. Method

In the following section we introduce our architecture, the different methods for merging the different modalities, our training and testing methods and the metrics we use to evaluate our system.

A. Proposed Architecture

Figure 1 shows our architecture, which consists of convolution (C), deconvolution (DC) and fully connected (N) layers. The network can be seen as consisting of five blocks, an image encoder, a label encoder, a merging layer, an image decoder, and a classification layer.

The architecture of the image encoder was a Convolutional Neural Network (CNN) based on [20], the decoder was a Deconvolutional Neural Network (DCNN) [21], created by inverting the CNN architecture and then adding two additional deconvolution layers to improve performance. The DCNN uses hyperbolic tangent (tanh) activation functions in all layers except the last which is a sigmoid.

The code for the model can be found at [22] and was implemented in python using the Keras deep learning library [23].

The multimodal architecture, in our case using two modalities, was trained using images and the class labels from the MNIST handwritten digit data set [14]. The network was tasked with image reconstruction and label prediction (see figure 1).

B. Multimodal Merging Layers

Different merging layers were tried to determine the best method for combining different modalities. These were *Concatenate* and *Add*. Concatenation is performed along axis 0 i.e. given the two (8, 1) vectors, a (16,1) vector is produced).

C. Training Regimes

Different training regimes were trialled to determine the best method for maximising label classification accuracy and image reconstruction fidelity. To measure reconstruction fidelity, Mean Squared Error (MSE) was used to calculate the loss between the reconstructed and target image, as well as the predicted label and target label.

The different training regimes were *Bimodal* and *Randomly Degraded*. For the *Bimodal* condition, all training inputs have both image and word vector data. *Randomly Degraded* replaces, at random, roughly one third of the image inputs and roughly one third of the word vector inputs (if an image input is removed, its word vector input cannot be removed, guaranteeing at least one modality is present). This results in a training set which is approximately one third bimodal, one

third images only and one third word vectors only. In line with the training methods employed in [4], the removed modality is replaced with zeros.

D. Testing Methods

Different testing regimes were utilised on all of the trained models. Testing was either, bimodal (images and labels), unimodal (images) or unimodal (labels). Testing using a single modality allows for the demonstration of whether anchoring of images and labels to the same representation has occurred. E.g. when only labels are provided, if the correct anchoring has occurred it is expected that the correct label will be predicted but also that the correct image will be produced. If the correct anchoring has not occurred, it is expected that only the correct label will be produced and an incorrect image will be produced. This would be seen as a high label accuracy and a large MSE with respect to the image. (and vice-versa for testing the anchoring of images).

All merging layers, training and testing types were four fold cross validated. The results reported are the mean values of the four trials for each configuration.

IV. Results

The best label accuracy for image only testing was achieved by the *Add* merging layer, 96.69%. The lowest image loss for label only testing was achieved by the *Concatenate* merging layer, 0.0539. The lowest total loss of 0.0337, was achieved by the *Concatenate* merging layer when trained in a fully bimodal manner with images and labels. Full results for all merging layer and training regimes can be found in table I. The loss (MSE) takes values from 0 to 1 with lower values being better and Accuracy takes a value from 0 to 1 with higher values being better.

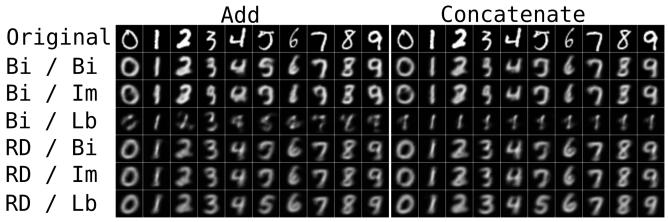

Fig. 2. Reconstructed Digits using different merging layers. Showing the original and reconstructed images where *Bi*, *RD*, *Im* and *Lb* are bimodal, randomly degraded, image and label data for training or testing.

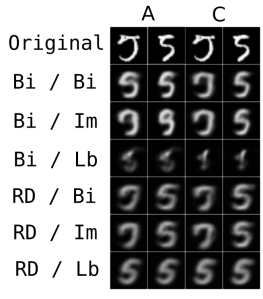

Fig. 3. Reconstructions of a poorly legible image of a five and a prototypical image of a five using different merging layers (*Concatenate* (C) and *Add* (A)) for different training and testing methods. Showing the original and reconstructed images where *Bi*, *RD*, *Im* and *Lb* are bimodal, randomly degraded, image and label data for training or testing.

V. Discussion

A. Effects of Different Training Regimes

Table I shows the results of *bimodal* and *randomly degraded* training for the different merge types.

For both merging layers, randomly degrading some of the training data lead to an improvement in label prediction accuracy when testing with only the images. This shows that when training with both modalities all the time, the network does not learn good image filters for the image encoder. This is likely due to the shallower nature of the label path compared to the image path, i.e. it is easier to ignore the image input and simply learn the label inputs in the fully bimodal situation.

Randomly degrading some of the data, forces the network to learn better filters for the image encoder. This is further highlighted in figure 2, where it can be seen that the reconstructed images for image only testing are better when randomly degraded training has taken place rather than fully bimodal training.

B. Effects of Different Merging Layers

Comparing the *Add* and *Concatenation* merging layers, we can see that whilst concatenation leads to the lowest total loss (table I), addition leads to a better mapping of word vectors to images in the bimodal condition and almost the same in randomly degraded training. In the bimodal condition this can be seen by the better reconstructed images in figure 2 for the *Bi/Lb* row as well as the lower image loss for *Add* versus *Concatenate*. The *Add* architecture also learns a better mapping of images to labels, as shown by the higher label prediction accuracy for image only testing.

C. Effects of Different Modalities

Looking at figure 2 it can be seen that the inclusion of label data helps improve the quality of image reconstruction for both merging types. This is also seen in the reduction of image loss.

In figure 3, the first example of an image of a five is interpreted as (possibly) an image of a nine or a seven by both merging layers when only visual information is present (rows *Bi/Im* and *RD/Im*. Looking at the reconstructed images for when only label information is present shows that both merging layers learnt a mapping for the label five and produce a prototypical 5 digit. Unfortunately combining both modalities did not greatly improve the quality of image reconstruction for the poorly written 5.

VI. Conclusion

We set out to learn a multimodal representation combining images and labels in order to anchor images and labels to the same abstract symbol which represents them i.e. anchoring an image of a one and the label one to the same latent representation. Two different methods for merging multimodal data were presented within the context of a deep multimodal autoencoder architecture.

Of these methods, addition (*Add*) gave the best quantitative results for both label accuracy as well as the subjectively best

TABLE I
LOSS AND ACCURACY OF DIFFERENT MERGING LAYERS FOR BIMODAL AND RANDOMLY DEGRADED TRAINING USING IMAGE AND WORD2VEC EMBEDDING DATA.

Merge Type	Training Data	Testing Data	Total Loss	Image Loss	Label Loss	Label Accuracy
Add	Bimodal	Bimodal	0.0345	0.0343	0.0002	1.0000
		Images	0.0942	0.0405	0.0537	0.6673
		Labels	0.0958	0.0799	0.0159	0.9268
	Randomly Degraded	Bimodal	0.0452	0.0451	0.0001	0.9997
		Images	0.0485	0.0434	**0.0051**	**0.9690**
		Labels	0.0541	0.0540	0.0002	1.0000
Concatenate	Bimodal	Bimodal	**0.0337**	0.0333	0.0003	0.9998
		Images	0.0945	0.0354	0.0591	0.6137
		Labels	0.1022	0.0903	0.0119	0.9266
	Randomly Degraded	Bimodal	0.0452	0.0451	0.0001	0.9994
		Images	0.0499	0.0437	0.0062	0.9634
		Labels	0.0542	**0.0539**	0.0003	1.0000

reconstructed images as seen in figure 2 (Particularly in the Bi/Lb condition).

VII. FUTURE WORK

Given the promising results obtained on this simple task, it will be interesting to see how this architecture can be extended to process more complex data as well as different modalities. In future the label input stream will be replaced with a sound input stream so that a system truly capable of unsupervised, life long learning can be produced, not having to rely on human annotations.

VIII. ACKNOWLEDGEMENTS

This research was funded by the EPSRC.

REFERENCES

[1] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[2] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

[3] L. W. Barsalou, "Grounded cognition," *Annu. Rev. Psychol.*, vol. 59, pp. 617–645, 2008.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

[5] M. Petit, S. Lallée, J.-D. Boucher, G. Pointeau, P. Cheminade, D. Ognibene, E. Chinellato, U. Pattacini, I. Gori, U. Martinez-Hernandez, *et al.*, "The coordinating role of language in real-time multimodal learning of cooperative tasks," *IEEE Transactions on Autonomous Mental Development*, vol. 5, no. 1, pp. 3–17, 2013.

[6] S. Fischer, D. Schulze, P. Borggrebe, M. Piefke, S. Wachsmuth, and K. Rohlfing, "Multi-modal anchoring in infants and artificial systems," 2011.

[7] R. A. Brooks, "Intelligence without representation," *Artificial Intelligence*, vol. 47, no. 1, pp. 139 – 159, 1991.

[8] R. Pfeifer and J. Bongard, *How the body shapes the way we think: a new view of intelligence*. MIT press, 2006.

[9] B. Graham, "Spatially-sparse convolutional neural networks," *CoRR*, vol. abs/1409.6070, 2014.

[10] G. M. Edelman, *Neural Darwinism: The theory of neuronal group selection.* Basic books, 1987.

[11] L. Smith and M. Gasser, "The development of embodied cognition: Six lessons from babies," *Artificial life*, vol. 11, no. 1-2, pp. 13–29, 2005.

[12] J. Piaget, "The origins of intelligence in children, new york (ww norton) 1963.," 1963.

[13] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 721–732, 2014.

[14] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, vol. 2, 2010.

[15] S. Coradeschi, A. Loutfi, and B. Wrede, "A short review of symbol grounding in robotic and intelligent systems," *KI-Künstliche Intelligenz*, vol. 27, no. 2, pp. 129–136, 2013.

[16] S. Coradeschi and A. Saffiotti, "Anchoring symbols to sensor data: preliminary report," in *AAAI/IAAI*, pp. 129–135, 2000.

[17] S. Coradeschi and A. Saffiotti, "An introduction to the anchoring problem," *Robotics and Autonomous Systems*, vol. 43, no. 2-3, pp. 85–96, 2003.

[18] A. G. Samuel, "Lexical activation produces potent phonemic percepts," *Cognitive psychology*, vol. 32, no. 2, pp. 97–127, 1997.

[19] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.

[20] Keras, "Mnist cnn keras github," *https://github.com/keras-team/keras/blob/master/examples/mnist_cnn.py*, 2017.

[21] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2528–2535, IEEE, 2010.

[22] E. Sheppard, "Bimodal autoencoder github," *https://github.com/ems7/simpleGrounders/blob/master/mnist/labels/backpropGrounder.py*, 2018.

[23] F. Chollet *et al.*, "Keras." https://github.com/keras-team/keras, 2015.