Deep Learning : A short overview

Nilani Algiriyage · Raj Prasanna · Kristin Stock · Emma E H Doyle · David Johnston

1 Deep Learning

AI is a field of study enabling machines to demonstrate the characteristics of human intelligence. Machine Learning (ML), a subset of AI, is a collection of algorithms that improve automatically through experience. They have been used to classify and cluster data to solve problems, such as spam detection, product recommendation, and online fraud detection, to name a few. Some of these ML algorithms need to be trained before they are used for research and are known as "supervised algorithms". To train the algorithms, the researcher must collect a large set of labelled data relevant to the problem. On the other hand, "unsupervised" algorithms do not require labelled training data [14]. There are also "semi-supervised" algorithms where learning is based on partially labelled data sets. DL is a subset of ML (see Fig. 1). The main difference between traditional ML and DL is how the features are extracted. Traditional ML approaches use handcrafted features by applying a variety of feature extraction algorithms and then apply learning algorithms. In contrast, the features are learned automatically by the DL algorithm and are interpreted hierarchically in multiple levels [14].

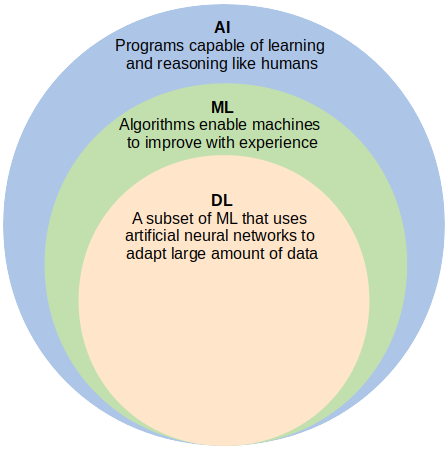

Fig. 1: The relationship between AI, ML and DL.

Nilani Algiriyage
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.
E-mail: r.nilani@massey.ac.nz

Raj Prasanna
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.
E-mail: r.prasanna@massey.ac.nz

Kristin Stock
Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand.
E-mail: k.stock@massey.ac.nz

Emma E H Doyle
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.
E-mail: e.e.hudson-doyle@massey.ac.nz

David Johnston
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.
E-mail: d.m.johnston@massey.ac.nz

The idea behind Artificial Neural Networks (ANNs), also known as Neural Networks (NNs), was inspired by the functioning of brain neurons. Generally, the brain can be represented as an interconnected set of nodes that can be organised in different layers; each layer generates outputs given certain inputs. DL algorithms are commonly identified under Deep Neural Networks (DNNs). A DNN is an ANN that has more than one layer of hidden nodes between its inputs and outputs (see Figures 2 and 3) [6]. DL algorithms can also be supervised: requiring labelled training data; unsupervised: not requiring labelled data for the training; or semi-supervised: partially requiring labelled training data.

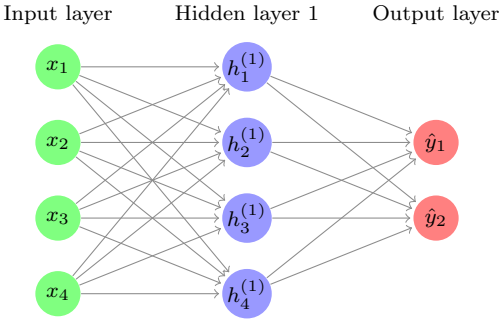

Fig. 2: Hidden layer of nodes in ANN.

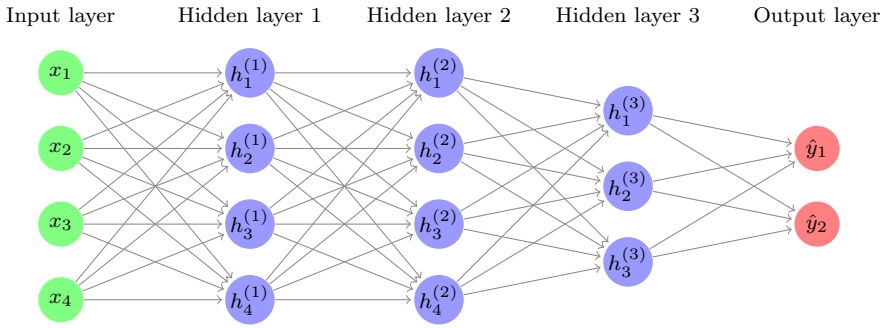

Fig. 3: Hidden layers of nodes in DNN.

As with the rapid increase in heterogeneous data sources and the multimodal data they provide, more sophisticated methods are required to analyse them. The main advantage of using DNNs is their ability to learn joint representations by correlating features in hidden layers. However, there exist different challenges in understanding multimodal representations, such as how to integrate data from heterogeneous sources, how to adapt to different levels of noise, and how to handle missing data [2]. The following section discusses the supervised DL techniques and the context of usage and their advantages and disadvantages.

1.0.1 Supervised Deep Learning Techniques

A supervised DL learning algorithm learns from the labelled training data and helps predict outcomes for unexpected data. We have identified Convolutional Neural Networks and Recurrent Neural Networks as more dominant supervised DL algorithms in many fields, including disaster research [19, 23].

Convolutional Neural Network (CNN): CNN, alternatively known as ConvNet, is a type of DNN, largely applied for object recognition in *computer vision* research. Typically, the layers that form the CNN architecture are known as the convolutional layer(s), the pooling layer(s), and the fully connected layer. CNNs are mainly used for image processing tasks and recognise the patterns across space (for example, they first recognise the lines and curves and then the full object in an image). Each convolutional layer operates a set of learnable parameters, called filters, that have smaller dimensions than the input image.

During the training process, these filters go through the whole input volume (for example, in the case of a colour image, the input volume consists of width * height * number of the RGB (red, green, and blue) channels, that is 3) and calculate an inner product of the input volume and the filter. This computation over the whole input leads to a feature map of the filter. The objective of the pooling layer is to progressively reduce the

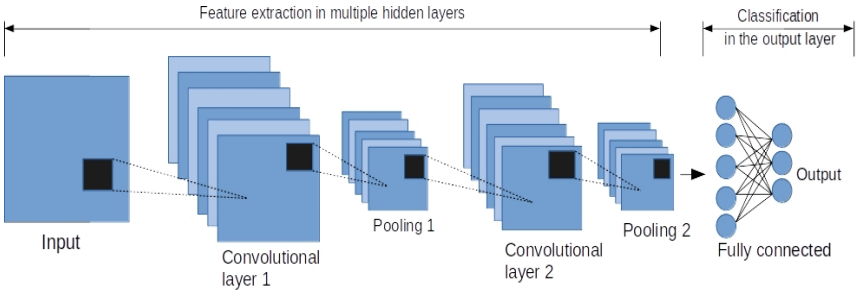

Fig. 4: The architecture of a CNN.

spatial size of the representation to reduce the number of parameters and computations in the network. The fully connected layer in the CNN represents the feature vector for the input. This feature vector is then further used for classification or translation to any other type of output [16, 18]. Over the years, variants of CNN architectures have been developed, such as AlexNet, DenseNet, VGGNet, and ResNet. More details of these architectures can be found in the survey paper by Khan et al. [11]. CNN has become very popular in multiple computer vision tasks, such as bidirectional images and sentence retrieval [10], emotion recognition [36], event recognition [32], and visual classification [16]. Fig. 4 shows the layout of a CNN.

Recurrent Neural Networks (RNN): RNNs are called recurrent, as they perform the same task for each element in a sequence. The input to an RNN consists of both the current sample and the previously observed sample. For example, the output of an RNN at time step $t-1$ affects the output at time step t. Each neuron is equipped with a feedback loop that returns the current output as an input for the next step. This structure can be expressed in such a way that each neuron in the RNN has an internal memory that keeps the information on the computations from the previous input [18].

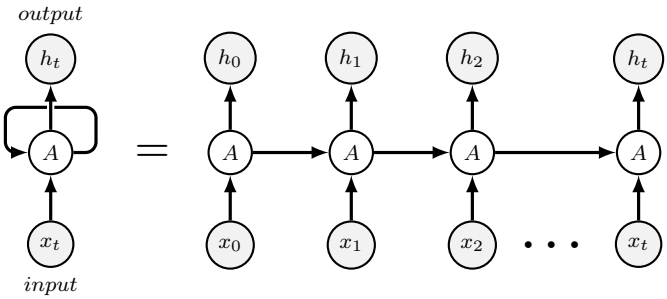

Fig. 5: The structure of a RNN.

The Long Short-Term Memory (LSTM) network is a variant of RNN. Both RNN and LSTM networks have been extensively used for analysing varying length sequences, such as videos, audio streams, and sentences [29, 30, 31, 34]. Fig. 5 depicts the structure of an RNN.

The following section discusses the unsupervised DL techniques and the context of usage, and their advantages and disadvantages.

1.0.2 Unsupervised Deep Learning Techniques

Unsupervised learning is a technique where we do not need to supervise the model. Instead, we allow the model to work on its own to discover patterns. Therefore, it mainly deals with the unlabelled data. Deep Belief Networks, Deep Boltzmann Machine, and Autoencoder are well-known unsupervised DL algorithms and are discussed below.

Deep Belief Networks (DBN): DBNs are a graphical representation that is essentially generative in nature as it produces all possible values that can be generated for the case at hand. DBNs consist of multiple layers with values. However, there is a relation between the layers but not the values. The main aim is to help the system classify the data into different categories. Srivastava et al. [25] introduced Multimodal Deep Belief Networks (DBN) to learn a joint density model over multimodal input space. In their multimodal DBM setting, two separate DBNs for text and image are trained in a completely unsupervised fashion and joined. Multimodal

DBNs have been applied in Audio-Visual Speech Recognition (AVSR) [8] and in gesture recognition [33], given the audio and skeleton features that assist the co-learning.

Deep Boltzmann Machine (DBM): DBM is a generative, stochastic model that can graphically represent as a set of interconnected visible and hidden nodes. Srivastava et al. [26] introduced a Multimodal-DBM. The applications of the algorithm include gesture recognition [33] and AVSR [8]. The key idea is to learn a distribution over multimodal inputs and fill the missing modalities using the conditional distribution of them given the observed distribution. The Multimodal-DBM model is capable of obtaining the joint representation even in the absence of some modalities, and the joint representation can imply real-world concepts. Also, it is possible to fill in the missing modalities given the observed ones due to their generative nature. DBMs have difficulty in training, high computational costs, and the need to use approximate variational training techniques [26]. The architectures of DBM and DBN are shown in Fig. 6.

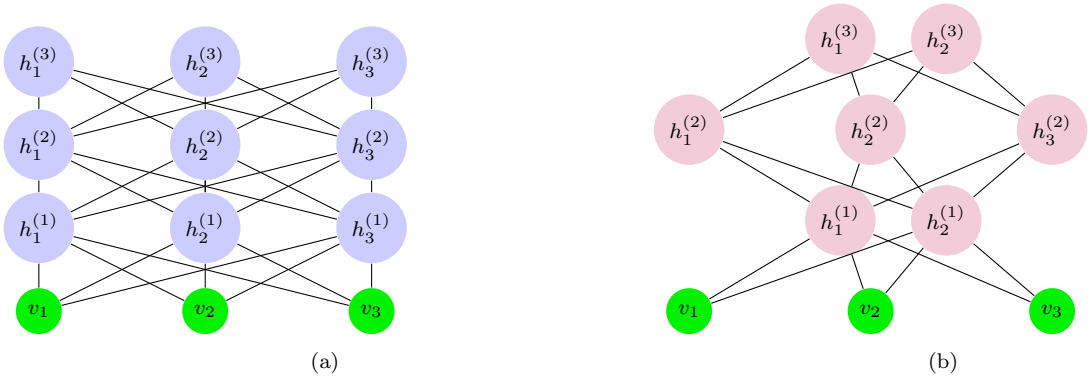

Fig. 6: The architectures of (a) DBM and (b) DBN.

Autoencoder (AE): Autoencoders (AEs) are a type of DNN, which have input, hidden, and output layers. However, the input layer is forced to be identical to the output layer. This network aims to reconstruct the input by transforming inputs into outputs in the simplest possible way such that it does not distort the input very much. AEs are used for dimensionality reduction. The work by Jiquan Ngiam et al. [21] introduced the use of AEs in a multimodal context. They used denoising autoencoders for each modality and fused them into a multimodal representation using another autoencoder layer. AEs have been successfully used in video retrieval [17], and video-based human pose recognition [7]. The main advantage is that the model does not need any labelled data for the pre-training. However, it cannot handle missing data. Fig. 7 illustrates the structure of a typical AE.

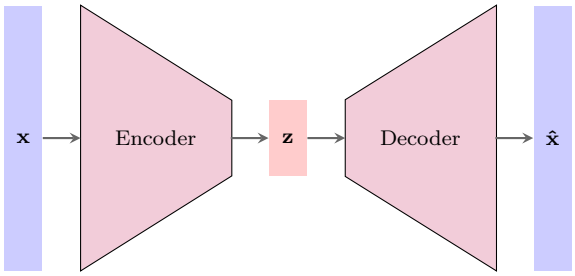

Fig. 7: The idea of an Autoencoder.

Table 1 summarises the supervised and unsupervised DL techniques and their applications. The survey paper by Zhang et al. [35] discusses more details of using DL models for BDA.

Table 1: Supervised and unsupervised DL techniques.

Application	Supervised DL Algorithm	Un-supervised DL Algorithm
1. Affect recognition • Emotion recognition • Personality trait recognition	CNN [36]	DBM [33], AE [7]
2. Event recognition • Human action & event recognition	CNN [32]	
3. Media description • Visual captioning • Visual Question Answering	RNN/LSTM [29]	
4. Multimedia retrieval • Bi-directional visual sentence search	CNN [10], RNN/LSTM	AE [17]
5. Speech recognition	CNN [27]	DBM , DBN [8]
6. Visual classification	CNN [16]	DBN

1.1 Semi-supervised Deep Learning Techniques

Semi-supervised deep learning is a class of DL algorithms that are able to learn from partially labelled data sets. Generative Adversarial Networks (GAN) and Domain-Adversarial Neural Networks (DANN) are used as a semi-supervised learning technique. Additionally, RNNs, including LSTM, are also used for semi-supervised learning.

Generative Adversarial Networks (GAN): GANs consist of two neural networks, namely the generative and discriminative networks, which work together to produce high-quality data [4]. The generator is responsible for producing new data after learning the data distribution from the training data set. The discriminator discriminates between actual data (coming from training data) and fake input data (coming from the generator). The objective function in GANs is based on minimax theory so that one network seeks to maximise the value function while the other network wants to minimise it. In each step, the generator, willing to fool the discriminator, produces sample data from random noise. The discriminator receives several real data examples from the training set along with samples from the generator. Then, the discriminator determines how good the generated samples are. The output of the discriminator helps the generator to optimise the generated data for the next round. The idea of a GAN is shown in Fig. 8. Having been inspired by GAN, Ganin et al.[3] proposed a Domain-Adversarial Neural Network (DANN) for semi-supervised problems that includes a component that explicitly aims to reduce the shift between a source and a target.

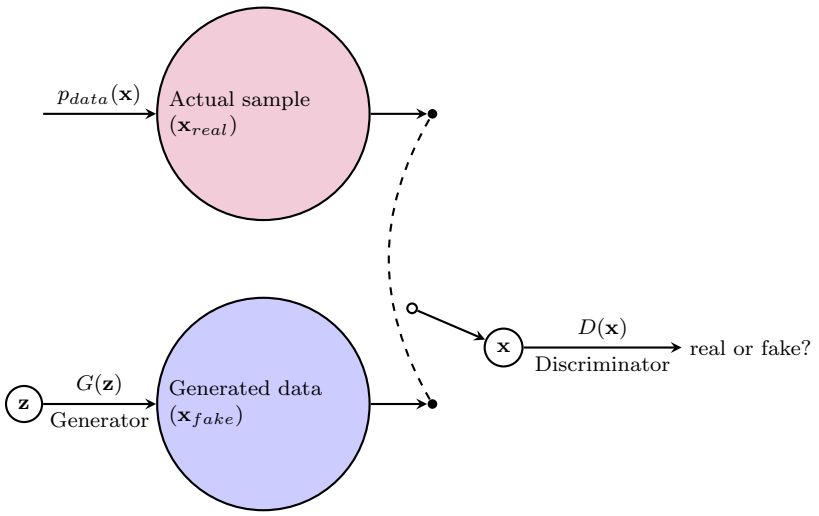

Fig. 8: The idea of a generative adversarial network.

.

It is a tedious task to train DL algorithms from scratch due to multiple reasons, such as dataset labelling and the computational power required for training. As a result, researchers have adopted multiple techniques such as transfer learning and domain adaptation to reuse already trained dataset.

1.2 Transfer Learning

Transfer learning is a highly adapted technique in DL research when there is insufficient data for a new domain. In transfer learning, a pre-trained DL model on a large data set (source) is applied to a new data set (target). The pre-trained network can be customised for 1) feature extraction and 2) fine-tuning to further train to make the model more relevant for the specific task. There are different sub-settings of transfer learning, such as inductive, transductive, and unsupervised transfer learning [24]. Table 2 shows the different transfer learning settings.

Table 2: Transfer learning settings.

Transfer learning setting	**Source Domain Labels**	**Target Domain Labels**
Inductive Transfer Learning	Available or unavailable	Available
Transductive Transfer Learning	Available	Unavailable
Unsupervised Transfer Learning	Unavailable	Unavailable

Domain Adaptation is a sub-class of transfer learning where data from a source domain is used to predict a target domain, under the assumption that the source and target domains have different distributions but share some similar patterns. It is related to transductive transfer learning, where the labels of the source domain data are available while the labels of the target domain data are not available. The task is to learn a classifier for the target data, using the labelled source data and the unlabelled target data. This technique has a high potential to be applied in the field of disaster research, given the lack of data just after a disaster [15]. Pan et al. [24] provide a comprehensive analysis of different transfer learning approaches. Table 3 shows the pre-trained DL models used in the papers we analysed.

Table 3: Transfer learning models used by the surveyed papers.

Author	Pre-trained Model	Data Modality	Description
[1, 13, 20, 22]	VGG-16	visual	A CNN trained on more than one million images from the ImageNet[1] database
[5, 15]	VGG-19	visual	A CNN trained on more than one million images from the ImageNet database
[9, 12]	GloVe	text	An unsupervised learning algorithm for obtaining vector representations for words
[28]	SoundNet	audio	A pre-trained model of natural sound representations using 2 million videos
[28]	Inception-v3	visual	A CNN trained on more than one million images from the ImageNet database

[1] http://www.image-net.org/

[19, 23]	AlexNet	visual	A CNN trained on more than one million images from the ImageNet database

References

1. Attari, N., Ofli, F., Awad, M., Lucas, J., Chawla, S.: Nazr-cnn: Fine-grained classification of uav imagery for damage assessment. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 50–59. IEEE (2017). doi:10.1109/DSAA.2017.72
2. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019)
3. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016). doi:10.1007/978-3-319-58347-1_10
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014)
5. Hezaveh, M.M., Kanan, C., Salvaggio, C.: Roof damage assessment using deep learning. In: 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 6403–6408. IEEE (2017). doi:10.1109/AIPR.2017.8457946
6. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al.: Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine **29** (2012)
7. Hong, C., Yu, J., Wan, J., Tao, D., Wang, M.: Multimodal deep autoencoder for human pose recovery. IEEE Transactions on Image Processing **24**(12), 5659–5670 (2015). doi:10.1109/TIP.2015.2487860
8. Huang, J., Kingsbury, B.: Audio-visual deep learning for noise robust speech recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7596–7599. IEEE (2013). doi:10.1109/ICASSP.2013.6639140
9. Kabir, M.Y., Madria, S.: A deep learning approach for tweet classification and rescue scheduling for effective disaster management. In: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 269–278 (2019). doi:10.1145/3347146.3359097
10. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems, pp. 1889–1897 (2014)
11. Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. arXiv preprint arXiv:1901.06032 (2019). doi:10.1007/s10462-020-09825-6
12. Kumar, A., Singh, J.P.: Location reference identification from tweets during emergencies: A deep learning approach. International journal of disaster risk reduction **33**, 365–375 (2019). doi:10.1016/j.ijdrr.2018.10.021
13. Kumar, A., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: A deep multi-modal neural network for informative twitter content classification during emergencies. Annals of Operations Research pp. 1–32 (2020). doi:10.1007/s10479-020-03514-x
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436 (2015). doi:10.1038/nature14539
15. Li, X., Caragea, D., Caragea, C., Imran, M., Ofli, F.: Identifying disaster damage images using a domain adaptation approach. In: Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Valencia, Spain. Academic Press (2019)
16. Li, Y., Ye, S., Bartoli, I.: Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning. Journal of Applied Remote Sensing **12**(4), 045008 (2018). doi:10.1117/1.JRS.12.045008
17. Liu, Y., Feng, X., Zhou, Z.: Multimodal video classification with stacked contractive autoencoders. Signal Processing **120**, 761–766 (2016). doi:10.1016/j.sigpro.2015.01.001
18. Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M.: Deep learning for iot big data and streaming analytics: A survey. IEEE Communications Surveys & Tutorials **20**(4), 2923–2960 (2018). doi:10.1109/COMST.2018.2844341
19. Muhammad, K., Ahmad, J., Baik, S.W.: Early fire detection using convolutional neural networks during surveillance for effective disaster management. Neurocomputing **288**, 30–42 (2018). doi:10.1016/j.neucom.2017.04.083
20. Naga Anitha, A.M.: Detection of disaster affected regions based on change detection using deep architecture. Procedia Computer Science **8** (2019)
21. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696 (2011)

22. Nguyen, D.T., Ofli, F., Imran, M., Mitra, P.: Damage assessment from social media imagery data during disasters. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 569–576 (2017). doi:10.1145/3110025.3110109
23. Pamuncak, A., Guo, W., Soliman Khaled, A., Laory, I.: Deep learning for bridge load capacity estimation in post-disaster and-conflict zones. Royal Society open science **6**(12), 190227 (2019). doi:10.1098/rsos.190227
24. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009). doi:10.1109/TKDE.2009.191
25. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: International conference on machine learning workshop, vol. 79 (2012)
26. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in neural information processing systems, pp. 2222–2230 (2012)
27. Tatulli, E., Hueber, T.: Feature extraction using multimodal convolutional neural networks for visual speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2971–2975. IEEE (2017). doi:10.1109/ICASSP.2017.7952701
28. Tian, H., Zheng, H.C., Chen, S.C.: Sequential deep learning for disaster-related video classification. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 106–111. IEEE (2018). doi:10.1109/MIPR.2018.00026
29. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014). doi:10.3115/v1/N15-1173
30. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164 (2015). doi:10.1109/CVPR.2015.7298935
31. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence **39**(4), 652–663 (2017). doi:10.1109/TPAMI.2016.2587640
32. Wang, L., Wang, Z., Du, W., Qiao, Y.: Object-scene convolutional neural networks for event recognition in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 30–35 (2015). doi:10.1109/CVPRW.2015.7301333
33. Wu, D., Shao, L.: Multimodal dynamic networks for gesture recognition. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 945–948. ACM (2014). doi:10.1145/2647868.2654969
34. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp. 2048–2057 (2015)
35. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Information Fusion **42**, 146–157 (2018). doi:10.1016/j.inffus.2017.10.006
36. Zhang, S., Zhang, S., Huang, T., Gao, W.: Multimodal deep convolutional neural network for audio-visual emotion recognition. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 281–284. ACM (2016). doi:10.1145/2911996.2912051