Deep Learning based Multimedia Analytics: A Survey

AUTHOR, College of William and Mary, USA

abstract goes here

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Artificial intelligence**; **Neural networks**;

Additional Key Words and Phrases: Multimedia Analytics, Deep Learning, Neural Networks

ACM Reference Format:
author. 2010. Deep Learning based Multimedia Analytics: A Survey. *ACM Trans. Web* 9, 4, Article 39 (March 2010), 4 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Todays digital contents are inherently multimedia: text, image, audio, video etc., due to the advancement of multimodal sensors. The multimedia data of interest cover a wide spectrum, ranging from text, audio, image, click-through log, Web videos to surveillance videos. Image and video contents, in particular, become a new way of communication among Internet users with the proliferation of sensor-rich mobile devices. Accelerated by tremendous increase in Internet bandwidth and storage space, multimedia data has been generated, published and spread explosively, becoming an indispensable part of todays big data.

Such large-scale multimedia data has opened challenges and opportunities for intelligent multimedia analysis, e.g., management, retrieval, recognition, categorization and visualization. Meanwhile, with the recent advances in deep learning techniques, we are now able to boost the intelligence of multimedia analysis significantly and initiate new research directions to analyze multimedia content. For instance, convolutional neural networks have demonstrated high capability in image and video recognition, while recurrent neural networks are widely exploited in modeling temporal dynamics in videos. Therefore, deep learning for intelligent multimedia analysis is becoming an emerging research area in the field of multimedia and computer vision.

The goal of this survey paper is to identify the typical scenarios and challenges emerging in multimedia analysis with deep learning techniques, review key tasks and the corresponding state-of-the-arts, introduce large scale real systems or applications, as well as discuss real-world datasets and benchmarks for future directions.

[5]

Author's address: author, College of William and Mary, 104 Jamestown Rd, Williamsburg, VA, 23185, USA, gang_zhou@wm.edu.

1559-1131/2010/3-ART39 $15.00
https://doi.org/0000001.0000001

2 FUNDAMENTAL TECHNIQUES

In this section, we first review basic building blocks of deep learning layer-wisely in Section 2.1, and then discuss several widely adopted network architectures in Section 2.2.

2.1 Basic Layers

Convolution: Convolution layer [5] applies a d-D[1] convolution over an input signal composed of several input channels. In most cases, *stride* and *padding* are adopted to adapt different input/output sizes.

Transposed convolution: Sometimes transposed convolution is also referred as fractionally-strided convolution or deconvolution, which goes in the opposite direction of a normal convolution. A convolution can be written as a product with sparse projection matrix $\mathbf{P}$ [13], while transposed convolution defined by $\mathbf{P}^T$ is widely adopted for upsampling, e.g., deblur [11], matting [14], super resolution [], image generation [9] and restoration [8].

Pooling layers combine multiple outputs at one layer into a single output in the next layer. **Max Pooling** and **Average Pooling** [10] are the most frequently adopted pooling layers that take the *max* or *average* on the input.

Normalization layers are essential in stabilizing training and accelerating convergence. **Batch Normalization** [4] reduces internal covariate shift in neural networks, which leads to faster convergence. Recent **Instance Normalization** [12] suits better for stylization task.

Activation layers are crucial since it introduces nonlinearity to the network. Traditional **Sigmoid** is best known for its continuity, recent **Relu**, **Noisy ReLU** [3] **Leaky Relu** [7], **ELU** [1] are all derived from rectified linear unit function, which have shown their superior performance in multimedia analysis [5].

Deep networks are versatile in defined various loss functions, ranging from **L1**, **MSE**, **CrossEntropy**, **Negative log likelihood** loss to **KL-divergence**, **Triplet Ranking** loss.

Dropout layer:

Sparse layer:

The most common initialization method includes Xavier initialization [2]

2.2 Network Architectures

2.2.1 Convolutional Neural Networks (CNN).

Convolutional neural network is first introduced decades ago [6]. Due to recent advancement of efficient computation hardware (GPUs) and abundant training data [**?**]. deep CNNs have recently shown an explosive popularity partially due to its success in image classification [18], [25]. They have also been successfully applied to other fields, such as object detection [33], [50], face recognition

2.2.2 Recurrent Neural Networks (RNN).

Besides the feed-forward network architecture, another big branch is based on the recurrent structures *RNN*, which suffers from the vanishing & exploding gradient problem. **LSTM** captures the long-term dependencies with the cell state. **GRU** further combines the forget and input gates into an update gate, and mergers the cell state and hidden state.

2.2.3 Generative Adversarial Networks.

[1]Typically, $d = 1$, 2, or 3 indicates 1D, 2D, 3D convolutions, respectively.

Fully-connected layer[2] connects all nodes in a layer to all nodes in its next layer.

3 HIGH-LEVEL TASKS

Detection, Recognition, xxx

tbd: level-wise? scope & coverage

4 BENCHMARKS

ilsvrc, cifar, coco, ...

REFERENCES

[1] Djork-Arn Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *CoRR* abs/1511.07289 (2015). http://dblp.uni-trier.de/db/journals/corr/corr1511.html#ClevertUH15

[2] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, Vol. 9. 249–256.

[3] Geoffrey E. Hinton. [n. d.]. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair. ([n. d.]).

[4] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, David Blei and Francis Bach (Eds.). JMLR Workshop and Conference Proceedings, 448–456. http://jmlr.org/proceedings/papers/v37/ioffe15.pdf

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105.

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* 1, 4 (Dec. 1989), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

[7] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

[8] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. 2016. Image Restoration Using Very Deep Fully Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. In *Advances in Neural Information Processing Systems (NIPS'16)*. arXiv:1603.09056 http://papers.nips.cc/paper/6172-image-restoration-using-very-deep-convolutional-encoder-decoder-networks-with-symmetric-skip-connections.pdf

[9] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *CoRR* abs/1511.06434 (2015). http://arxiv.org/abs/1511.06434

[10] Dominik Scherer, Andreas Müller, and Sven Behnke. 2010. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In *Proceedings of the 20th International Conference on Artificial Neural Networks: Part III (ICANN'10)*. Springer-Verlag, Berlin, Heidelberg, 92–101. http://dl.acm.org/citation.cfm?id=1886436.1886447

[11] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. 2016. Deep Video Deblurring. *arXiv preprint arXiv:1611.08387* (2016).

[12] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR* abs/1607.08022 (2016).

[13] Dumoulin Vincent and Visin Francesco. 2016. A guide to convolution arithmetic for deep learning. In *arXiv:1603.07285*.

[14] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. 2017. Deep Image Matting. *CVPR* (2017).

[2] Also known as linear layer

Received February 2007; revised March 2009; accepted June 2009