A Review on Deep Learning based Multimedia Analytics

Multimedia analytics is of great interests for multimedia researchers. At the same time, multimedia community has also witnessed the rise of deep learning based techniques in analyzing multimedia content more effectively ...

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:
. 2010. A Review on Deep Learning based Multimedia Analytics. *ACM Trans. Web* 9, 4, Article 39 (March 2010), 6 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Due to the advancement of multimodal sensors, today's digital contents are inherently multimedia, e.g., text, image, audio, video. The multimedia data of interest covers a wide spectrum, ranging from text, audio, image, click-through log, Web videos to surveillance videos. Visual content, i.e., images and videos, 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 today's big data.

Such large-scale multimedia data has opened challenges and opportunities for intelligent multimedia analysis, e.g., management, retrieval, recognition, categorization, visualization and generation. Meanwhile, with 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; recurrent neural networks are widely exploited in modeling temporal dynamics in videos; and generative adversarial networks are capable of generating realistic images on demand. 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 review state-of-the-arts deep learning components and network architectures, to identify typical scenarios and challenges emerging in multimedia analysis, and to discuss real-world datasets and benchmarks for future directions.

Author's address:

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

2 PRELIMINARY

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 Building blocks

Most popular deep learning frameworks are highly modularize, such that deep networks can be easily constructed by a collection of interacting layers.

Convolution layer [15] applies the convolution operation over an input signal, which is especially critical for multimedia visual data. Transposed convolution[1] goes in the opposite direction [34], and is widely adopted for upsampling in deblurring [32], image matting [36], super resolution [16], image generation [25] and restoration [19].

Fully connected layer [27] defines a linear transformation between layers, where neurons in one layer is connected to all neurons in another layer. A typical functionality is high-level reasoning [15] after several convolution layers.

Activation [2, 7, 8, 18, 21] and **Pooling layers** [15, 31] introduce nonlinearity into networks, which have demonstrated their superior performance in multimedia analytics. In principle, activation defines the response mechanism for neuron outputs, and pooling combine multiple outputs at one layer into a single output in the next layer, which introduces a form of non-linear down-sampling.

Normalization layers [10, 33] are critical in stabilizing the training process and accelerating the convergence speed. For example, Batch Normalization [10] reduces internal covariate shift in neural networks, which leads to faster convergence.

Loss layers define various loss functions for diverse purposes, ranging from *L1*, *MSE*, *Cross Entropy*, *Negative log likelihood* losses to *KL-divergence* and *Triplet Margin* losses.

Optimizing deep networks is generally difficult. On one hand, different initializations [4, 9] have major impact on network convergence. On the other hand, the optimization algorithm is also important for convergence. The fundamental optimization technique is based on back-propagation [28] via SGD [24, 30], Adam [14], LBFGS, Rprop, or RMSprop [6].

2.2 Network Architectures

2.2.1 *Convolutional Neural Networks (CNN).*

Convolutional neural network is first introduced decades ago [17] for recognizing zip codes, which is rather primitive at that time. Recent advancements in computation hardware (GPU[2]) and abundant training data [29] bring prosperity to CNN architecture [15].

...

2.2.2 *Recurrent Neural Networks (RNN).*

Besides the feed-forward network architecture, another big branch is based on the recurrent structures *RNN*. However, RNN suffers from the vanishing & exploding gradient problem since it was invented. **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]In some literatures, it is also referred as fractionally-strided convolution or deconvolution.

[2]Graphics Processing Unit

Generative Adversarial Networks (GAN) [5] gains great attention since it only defines the high-level objective (real/fake) rather than a fixed loss function.

In a short time period, many variants, e.g., wGAN[1], DCGAN[25], have improved the original framework for generating more realistic images robustly. Laplacian Pyramid of Adversarial Networks [3] extends GAN for progressively generating images with higher resolution. More recent work by NVIDIA [12] futher generates celebrity photos in impressive quality. Conditional GAN [20] takes extra input for generating images based on a constraint. Pix2pix [11] translates an image to another representation with conditional GAN. CycleGAN [39], DiscoGan [13] and DualGan [37] share the same idea for image translation between different domains, where unpaired data is adopt in a self-supervised way. Moreover, conditional GAN is also applied for generating images conditioning on text descriptions [23, 26, 38] of bird [35] or flower [22].

3 HIGH-LEVEL TASKS

This section reviews high-level tasks in multimedia analytics built upon deep learning framework. This survey covers various task involved in multimedia analytics, with a focus on visual data, i.e., image and video.

Fig. 1 shows the roadmap of multimedia analytics, which focuses on analyzing multimedia data (left) from diverse perspectives (right).

- multimedia → label.
- multimedia → location.
- multimedia → sentence.
- multimedia → paragraph.
- multimedia → image.
- Image/Video generation (reverse mapping). Generating images given a label, paragraph texts, etc..

For example, image classification aims to translate an image to one or multiple labels, while image captioning parse an input image to a sentence. The difficulty of each task increases as the capacity of target domain grows. Moreover, the reverse mapping (from left to right) is also valid, due to recent advancement of Generative Adversarial Networks (GAN).

3.1 Detection

Fig. 2 summarizes several milestones for image classification on ILSVRC dataset.

3.2 Localization

3.3 Captioning

3.4 Generation

4 BENCHMARKS

This section reviews several popular benchmarks on multimedia analytics, as well as state-of-the-arts advancements on their benchmarks.

ilsvrc,

cifar,

coco,

...

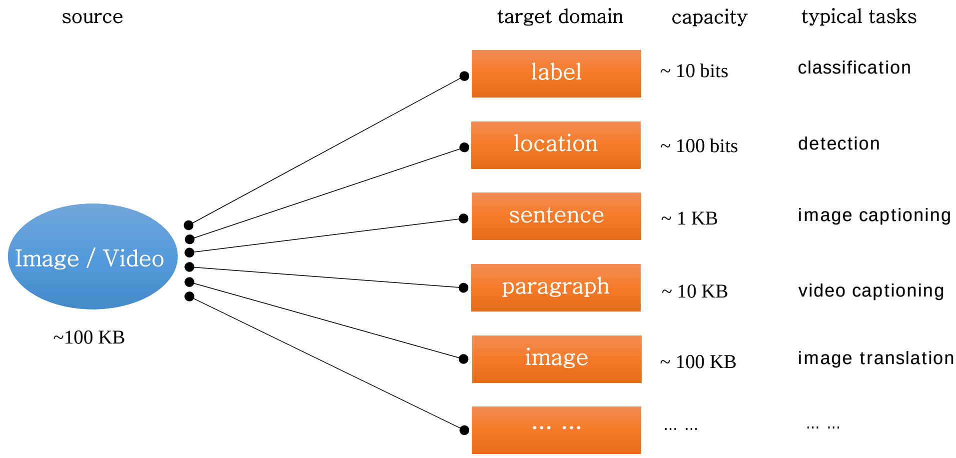

Fig. 1. A road-map for deep learning based multimedia analytics.

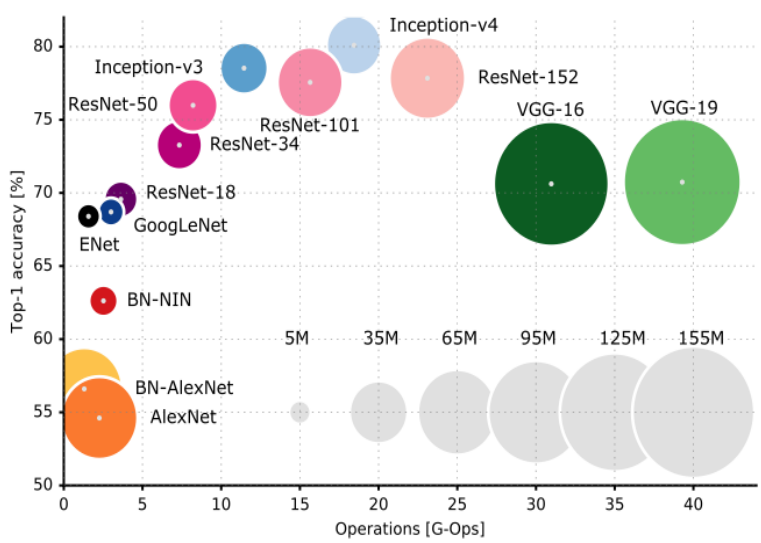

Fig. 2. Major milestone on ILSVRC.

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou. 2017. Wasserstein GAN. In *arXiv:1701.07875*.
[2] 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).
[3] Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in Neural Information Processing Systems 28*. 1486–1494.

[4] 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.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[6] Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *CoRR* abs/1308.0850 (2013).

[7] Richard H. R. Hahnloser, Rahul Sarpeshkar, Misha Mahowald, Rodney J. Douglas, and H. Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405 6789 (2000), 947–51.

[8] Jun Han and Claudio Moraga. 1995. *The influence of the sigmoid function parameters on the speed of backpropagation learning*. 195–201.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. 1026–1034.

[10] 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)*. JMLR Workshop and Conference Proceedings, 448–456.

[11] P. Isola, J. Zhu, and A. A. Efros T. Zhou. 2016. Image-to-Image Translation with Conditional Adversarial Networks. In *arXiv:1611.07004*.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *arXiv:1710.10196v2*.

[13] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *ICML*.

[14] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*. 1097–1105.

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *arXiv:1704.03915*.

[17] 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.

[18] 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*.

[19] 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)*.

[20] M. Mirza and S. Osindero. 2014. Conditional Generative Adversarial Nets. In *arXiv:1411.1784*.

[21] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)*. 807–814.

[22] M-E. Nilsback and A. Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing.*

[23] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2016. Conditional Image Synthesis With Auxiliary Classifier GANs. *arXiv* (2016). https://arxiv.org/abs/1610.09585

[24] B. T. Polyak and A. B. Juditsky. 1992. Acceleration of Stochastic Approximation by Averaging. *SIAM J. Control Optim.* 30, 4 (July 1992), 838–855.

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

[26] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text-to-Image Synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. Chapter Learning Internal Representations by Error

Propagation, 318–362.
[28] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1988. Neurocomputing: Foundations of Research. MIT Press, Cambridge, MA, USA, Chapter Learning Representations by Back-propagating Errors, 696–699.
[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.
[30] David Saad (Ed.). 1998. *On-line Learning in Neural Networks.* Cambridge University Press, New York, NY, USA.
[31] 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).* 92–101.
[32] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. 2016. Deep Video Deblurring. *arXiv preprint arXiv:1611.08387* (2016).
[33] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR* abs/1607.08022 (2016).
[34] Dumoulin Vincent and Visin Francesco. 2016. A guide to convolution arithmetic for deep learning. In *arXiv:1603.07285.*
[35] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. 2010. *Caltech-UCSD Birds 200.* Technical Report CNS-TR-2010-001. California Institute of Technology.
[36] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. 2017. Deep Image Matting. *CVPR* (2017).
[37] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. (2017).
[38] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. 2017. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In *arXiv:1612.03242.*
[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV.*

Received February 2007; revised March 2009; accepted June 2009