Towards Neural-Symbolic AI for Media Understanding

Polyana B. Costa
polyana@telemidia.puc-rio.br
PUC-Rio, Brazil

Guilherme Marques
guilherme.marques@telemidia.puc-rio.br
PUC-Rio, Brazil

Arhur C. Serra
arthursrr@telemidia.puc-rio.br
PUC-Rio, Brazil

Daniel de S. Moraes
danielmoraes@telemidia.puc-rio.br
PUC-Rio, Brazil

Antonio J. G. Busson
busson@telemidia.puc-rio.br
PUC-Rio, Brazil

Álan L. V. Guedes
alan@telemidia.puc-rio.br
PUC-Rio, Brazil

Guilherme Lima
guilherme.lima@ibm.com
IBM Research, Brazil

Sérgio Colcher
colcher@inf.puc-rio.br
PUC-Rio, Brazil

ABSTRACT

Methods based on Machine Learning have become state-of-the-art in various segments of computing, especially in the fields of computer vision, speech recognition, and natural language processing. Such methods, however, generally work best when applied to specific tasks in specific domains where large training datasets are available. This paper presents an overview of the state-of-the-art in the area of *Deep Learning* for Multimedia Content Analysis (image, audio, and video), and describe recent works that propose The integration of deep learning with symbolic AI reasoning. We draw a picture of the future by discussing envisaged use cases that address media understanding gaps which can be solved by the integration of machine learning and symbolic AI, the so-called Neuro-Symbolic integration.

KEYWORDS

neural-symbolic computing, media understanding, deep learning, reasoning

1 INTRODUCTION

Methods based on Machine Learning, in particular Deep Learning, have become state-of-the-art in various segments of computing, especially in the fields of computer vision, speech recognition, and natural language processing [3]. However, such methods generally work best when applied to specific tasks in domains where large training datasets are available. Ask a voice assistant something more complex than command and it will probably struggle. Recently, researchers began to investigate a new paradigm called Neuro-Symbolic AI aiming for the development of machines with "common sense" [30]. It tries to integrate methods from both Machine Learning (Neuro part) and Symbolic Reasoning (Symbolic part) fields.

According to Valiant [35], one of the key challenges for AI is the construction of integrated machine learning and reasoning mechanisms that can exhibit human-like cognitive behavior. In this way, researchers want AI to not only recognize objects or voice commands but to also be able to understand what it sees/hears and apply reasoning to act accordingly. Figure 1 shows two examples of image captions generated by Deep Learning. Neural networks can correctly identify the key objects in a scene but they cannot completely understand how these objects relate to each other to produce the correct description of the image [22]. Such a problem is addressed by a neuro-symbolic approach from Maio *et. al.* [26], called Concept Learner (NS-CL). It learns visual concepts, words, and semantic parsing of sentences while it builds an object-based scene representation and translates sentences into symbolic programs, allowing questions and answering about the elements of the scene.

Figure 1: Image captions are generated by Deep Learning [22]

In this position paper, we present an overview of the state-of-the-art in the field of Deep Learning for Multimedia Content Analysis (image, audio, and video), and describe recent works that propose their integration with more traditional forms of symbolic reasoning. In particular, we cite the Deep Reasoning Networks (DRNets), which combine both deep learning and reasoning to infer. They use logic and constraint reasoning to exploit the structure of a problem and extract prior knowledge. Then, we draw a picture of future vision by describing envisaged use cases that address media understanding gaps that can be solved by using neuro-symbolic methods.

The rest of this article is organized as follows. Section 2 describes *state-of-the-art* in both Multimedia Analysis using Deep Learning and neuro-symbolic methods. section 3 presents our future vision regarding Media Understating. Finally, section 4 contains the final considerations.

2 STATE OF ART

2.1 Multimedia Content Analysis

2.1.1 Image Classification. In multimedia, the classification task consists of mapping media content into one or more distinct categories. Deep Leaning architectures based on CNN's (Convolutional neural network) or ConvNets have become the main method used for recognizing audiovisual patterns. Typically the training of CNNs is done in a supervised manner, and they are trained over datasets, which contains thousands/millions of media and related classes.

In: VII Workshop "O Futuro da Videocolaboração" (WCT-Video 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.

ISSN 2596-1683

During training, CNNs learn the hierarchy of *features* that are applied to the input media so that it is possible to classify their content. Since the AlexNet [20] victory in the ImageNet 2012 challenge [31], new and more accurate CNN-based architectures have emerged. The SE-Net architecture (*Squeeze-and-Excitation Network*) [15] is *state-of-the-art* in the image classification task, obtaining 2.25% error top-5 at ImageNet 2017. SE-Net proposes a new type of block called SE, which improves the network's power of representation by highlighting the interdependencies between the image channels and their *features* maps. For this, SE-Net uses a mechanism that allows the network to recalibrate *features*, through which it uses global information to emphasize the most informative *features* and suppress the *features* less useful.

2.1.2 Video Classification. Unlike images, videos have not only visual, but also contain auditory information. Current methods for classifying video generally consist of two stages: (1) CNNs, called *backbones*, are used to extract the audio-visual *features* from the video content; (2) After extraction, sophisticated methods for aggregating *features* such as NetVLAD [1] and NetFV [29] can be applied to undermine audio-visual *features* and perform classification. These methods achieve state-of-the-art performance on the YouTube-8M video classification task [32]. Currently, the video analysis community is devoted to classifying video at the segment level (i.e., temporally locate and classify the segments in the videos). Deep recurrent models such as LSTMs [14] and GRUs are commonly used for video segment classification as they are well suited to extract temporal features across time.

2.1.3 Action Recognition. Methods for action recognition on trimmed videos have achieved impressive results. For instance, [10] proposed a two-pathway model called SlowFast where one slow pathway is designed to capture static, spatial information and, a second fast pathway responsible for capturing motion, dynamic information. This method achieves state-of-the-art accuracy on the Kinectis-400 dataset [19] benchmark without any extra training data. However, in realistic problems, videos do not come temporally trimmed. Because of that, researchers have been giving increasing attention to the task of temporal action localization (TAL), which aims to predict the temporal boundary and label of actions in untrimmed videos. This is a very challenging setting because datasets for this task come with videos that may have a very long duration. Recently, self-supervised learning has caught the attention of the computer vision community due to its ability to learn informative features without human supervision. It works by designing an auxiliary task that labels can be self-annotated. For example, [5] proposed an auxiliary task that predicts temporal permutation for cross-domain videos to tackle the problem of Spatio-temporal variations for action segmentation. This self-supervised approach combined with MS-TCN has improved the MS-TCN stand-alone version accuracy on all three datasets 50Salads [33], GTEA [9] and Breakfast [21], and requires only 65% of the labeled training data for comparable performance.

2.1.4 Sentiment Analysis. The area of sentiment analysis is large and comprises many subtopics, such as the study of humans' emotions regarding conversations, events, or in general [25]. It is a large area with many subtopics, as social media sentiment analysis, emotion recognition, polarity classification, etc. The multimodal data of a video consists of frames, audio, and audio transcript content. A dataset like IEMOCAP [2] contains all of these annotated data modules which enable a more consistent sentiment analysis in a video. Currently, the state-of-the-art for this dataset consists of extracting the transcription features through a CNN [18], the audio features through the openSMILE software [8], and the visual features through a 3D-CNN [17]. Using a bi-directional contextual LSTM (bc-LSTM) it is possible to convert each feature into a single format, concatenating and classifying them with another bc-LSTM. Reaching 76.1% accuracy of classification. Considering only the audio data, [7] reaches 66.9% weighted accuracy, using a CNN+LSTM approach. Finally, considering only the transcribed content, [23] reaches 60.84% accuracy using a strategy similar to a bi-directional LSTM.

Human emotions can be expressed through verbal or body language. Both may simulated though human intentionally acting, when try to suppress them or demonstrate another emotions [27]. In micro-expressions, mostly facial muscle movements contains a significant and effective amount of information about the true emotions felt at the moment of manifestation. Seeking the consistency of a sentiment analysis through micro-expressions, some datasets have already been produced that map universal emotions (happy, sad, anger, fear, surprise, disgust, and contempt) into spontaneous micro-expressions through video content [28]. Some spontaneous micro-expressions datasets include: SMIC [24], CASME [36], CASME II [38], and SAMM [6]. Only in this context, we can unite devout lines of research in computer vision, such as face detection, facial point mapping, temporal series classification, and semantic summarization.

2.2 Neuro-Symbolic Models

A few decades ago [13] started to discuss how the integration of both symbolic and connectionist AI could suppress the limitations of each paradigm. Some of these works [34] emphasized that connectionist approaches had a hard time capturing high-level cognitive processes, because most of these processes are structured and systematic, which contrasts with the characteristics of conventional neural network models. Moreover, symbolic models are capable of capturing a wide variety of intelligence and making inferences about the obtained information. However, the connectionist approach manages to obtain knowledge from training examples, due to its ability to generalize and represent complex and inaccurate knowledge [12]. The combination of these approaches aims to integrate two fundamental cognitive skills: the ability to learn from the environment and to reason over data [11]. This unified approach is called neuro-symbolic AI.

More recently, advances in neuro-symbolic AI have produced models capable of solving complex tasks. For example, Deep Reasoning Networks (DRNets) combine both deep learning and reasoning to infer crystal structures of materials from X-ray diffraction data under thermodynamic rules [4]. Along with an stochastic-gradient-based neural network optimization, DRNets use logic and constraint reasoning to exploit the structure of a problem and extract prior knowledge. In another unified approach, the Neuro-Symbolic Concept Learner (NS-CL) learns visual concepts, words, and semantic parsing of sentences [26]. NS-CL builds an object-based scene

representation and translates sentences into symbolic programs, allowing question and answering about the elements of the scene. Another work used a graph-based network module to detect action in videos, by reasoning over the temporal relations present in each video [16].

Despite the advances achieved by these models, some studies showed that most neuro-symbolic algorithms still struggle with understanding causal relations over time. For instance, Clevrer [39] cites that such algorithms perform poorly when they have to provide explanatory and predictive information. Such information refers to why something happened, and not only to notify that it did *e.g.* the actors are cold because previously they got out in the rain. These limitations show that there are still research gaps to address in favor of improving neuro-symbolic models.

3 FUTURE VISION

We present our future vision through six envisaged use cases

UC-01 Media Interpretation. Media classification is a problem that has been extensively investigated for decades. Currently, benchmark models can categorize images, audio, and video into thousands of classes. However, machine learning alone is not suitable for scenarios where high-level human interpretation is needed, especially in content that involves irony, such as montages and internet MEMEs. Figure 2 shows two antagonistic examples of internet MEMES about Donald Trump used in Twitter during the U.S. presidential election in 2016. Note that this type of content is common on the internet and is a common communication artifact among young people. Developing an AI model to interpret this type of content automatically involves recognizing entities (people, objects, etc.) present in the media and understanding how these recognized entities are related and what these relations mean.

Figure 2: Antagonistic examples of internet MEMES about Donald Trump used in Twitter during the U.S. presidential election in 2016.

UC-02 Content-Aware Applications. Understanding the meaning of media also opens up the opportunity for multimedia application development that explore the semantics of the content to perform synchronism. Imagine, for example, an intelligent advertising system that triggers an advertisement for a soccer brand when the audience watches a video clip that is in the context of interest of the brand (*e.g.* when the video's presenter comments on soccer, or when a section of a person scoring a goal appears in a soccer game). The complexity of such an application goes beyond simple video classification or action recognition, as it requires a high degree of understanding about the context of the video scene.

UC-03 Semantic Summarizing. The objective of video summarization is to create a compact representation of a given video. Recently, researchers have been focusing on the task of query-focused video summarization. The work of [37] proposes a method called Convolutional Hierarchical Attention Network (CHAN) that selects important visual features and computes the similarity of them with a given query. Despite interesting results, this approach is not able to reason about high-level, contextual concepts present in queries, nor logical, temporal, or causal structures. Future systems with neuro-symbolic capabilities will be able to perform summarizations based on queries that understand complex reasoning tasks such as to cause and effect, entities, and relationships. For instance, given a video of a soccer match where player X is kicked out of game after repeated faults, a query such as "Summarize the reason player X was sent off" would demand a system that is capable of determining that the reason was because that player X received a red card and performed multiple faults in the game.

UC-04 Sentiment Context Interpretation. One way to go beyond the simple metadata generation of emotion classification is the semantics of media data (video, image, music). By knowing the context of data that generates a sentimental reaction, and being able to catalog them, we can learn how media data induce feelings in a person. For example, by knowning the contexts or elements in a film or publicity campaign that produces a certain emotion, use them to generate the desired effect on the target audience. The same method may also be used in conversation in call centers. It is possible to learn what attendant support actions and offered services will be a positive effect on consumers. A service that can signal to the attendant which behaviors or actions can lead the client to some emotional tension, considering that the attendant (person or bot) have no way to analyze these nuances of behavior.

UC-05 Semantic Query in Video Classes. Machine Learning tasks in video classes may capture a lot of audio-visual information such as who is the lecturer and labels considering his(er) speaking and used graphics. For instance, a video in which a professor is drawing equations can be tagged as Math class. In addition, the integration of the knowledge data related to teaching topics can support semantic queries over education videos, especially by relating video moments. For instance, it may help to locate the moment that a professor shows mathematical proof of an specific equation, e.g., Pythagora's theorem, and from this link related information. Or it may recommend videos from the same professor in which topics presented, in a previous video, continue.

UC-06 Music Understanding. Some music lyrics require background knowledge to be better understood. This knowledge is strongly related to the genre and authors. Moreover, the music melody and rhythm can also have knowledge related to an author, a genre. The neuro-symbolic can help to integrate both pieces of knowledge and lead to a better understanding of music genres and improve search/recommendation tasks. For instance, in "Garota de Ipanema" song from Tom Jobim, the calm and simple rhythm of the "Bossa Nova" genre may be related to the bohemian and beachy attitute, popular in the Rio de Janeiro, the city where the Ipanema beach from the song is located. Another scenario would be to find references (through connected knowledge) for some authors in other music, to a better grouping or navigation through music with

the same rhythm, similar melodies, same authors, and are address similar topics.

4 FINAL REMARKS

In this paper, we presented an overview of the state-of-the-art in the area of *Deep Learning* for multimedia content analysis (image, audio, and video) and described recent works that propose the integration of machine learning and Neuro-Symbolic methods. Then we draw a picture of the future regarding the Media Understating by list a set of envisaged use cases. Our main research goal is exactly how enable Machine Learning use the structured knowledge bases and linked data that already exist to be able to solve problems that ML alone cannot solve.

The work is part of ongoing research that aims to improve Media Understanding by using recent Neuro-Symbolic AI. We mention as future evaluate the usage of such methods in FakeNews and DeepFakes datasets.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5297–5307.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[3] Antonio José G Busson, Lucas Caracas de Figueiredo, Gabriel NP dos Santos, André Luiz de B. Damasceno, Sérgio Colcher, and Ruy Milidiú. 2018. Developing Deep Learning Models for Multimedia Applications in TensorFlow. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. 7–9.

[4] Di Chen, Yiwei Bai, Wenting Zhao, Sebastian Ament, John M Gregoire, and Carla P Gomes. 2019. Deep reasoning networks: Thinking fast and slow. *arXiv preprint arXiv:1906.00855* (2019).

[5] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. 2020. Action Segmentation With Joint Self-Supervised Temporal Domain Adaptation. 9454–9463. https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Action_Segmentation_With_Joint_Self-Supervised_Temporal_Domain_Adaptation_CVPR_2020_paper.html

[6] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2016. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing* 9, 1 (2016), 116–129.

[7] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. 2018. Cnn+ lstm architecture for speech emotion recognition with data augmentation. *arXiv preprint arXiv:1802.05630* (2018).

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[9] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. 2011. Learning to recognize objects in egocentric activities. In *CVPR 2011*. 3281–3288. https://doi.org/10.1109/CVPR.2011.5995444 ISSN: 1063-6919.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. 6202–6211.

[11] Artur d'Avila Garcez, Marco Gori, Luis C Lamb, Luciano Serafini, Michael Spranger, and Son N Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088* (2019).

[12] Ioannis Hatzilygeroudis and Jim Prentzas. 2010. Integrated rule-based learning and inference. *IEEE Transactions on Knowledge and Data Engineering* 22, 11 (2010), 1549–1562.

[13] Mélanie Hilario, Yannick Lallement, Frédéric Alexandre, and Crin-inria Lorraine. 1995. Neurosymbolic integration: Unified versus hybrid approaches. In *In The European Symposium On Artificial Neural Networks*. Citeseer.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997).

[15] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-Excitation Networks. *CoRR* abs/1709.01507 (2017). arXiv:1709.01507 http://arxiv.org/abs/1709.01507

[16] Yifei Huang, Yusuke Sugano, and Yoichi Sato. 2020. Improving Action Segmentation via Graph-Based Temporal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14024–14034.

[17] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 221–231.

[18] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *arXiv:1705.06950 [cs]* (May 2017). http://arxiv.org/abs/1705.06950 arXiv: 1705.06950.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[21] Hilde Kuehne, Ali Arslan, and Thomas Serre. 2014. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, 780–787. https://doi.org/10.1109/CVPR.2014.105

[22] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40 (2017).

[23] Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2020. BiERU: Bidirectional Emotional Recurrent Unit for Conversational Sentiment Analysis. *arXiv preprint arXiv:2006.00492* (2020).

[24] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen. 2013. A Spontaneous Micro-expression Database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–6.

[25] S Maghilnan and M Rajesh Kumar. 2017. Sentiment analysis on speaker specific speech data. In *2017 International Conference on Intelligent Computing and Control (I2C2)*. IEEE, 1–5.

[26] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584* (2019).

[27] David Matsumoto, Seung Hee Yoo, and Sanae Nakagawa. 2008. Culture, emotion regulation, and adjustment. *Journal of personality and social psychology* 94, 6 (2008), 925.

[28] Walied Merghani, Adrian K Davison, and Moi Hoon Yap. 2018. A review on facial micro-expressions analysis: datasets, features and metrics. *arXiv preprint arXiv:1805.02397* (2018).

[29] Antoine Miech, Ivan Laptev, and Josef Sivic. 2017. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905* (2017).

[30] Katia Moskvitch. 2020. Neurosymbolic AI to Give Us Machines With True Common Sense. https://medium.com/swlh/neurosymbolic-ai-to-give-us-machines-with-true-common-sense-9c133b78ab13

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015).

[32] Miha Skalic and David Austin. 2018. Building a size constrained predictive model for video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.

[33] Sebastian Stein and Stephen J. McKenna. 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 729–738. https://doi.org/10.1145/2493432.2493482

[34] Ron Sun. 1993. On neural networks for symbolic processing. In *Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*. IEEE, 5–6.

[35] Leslie G. Valiant. 2003. Three Problems in Computer Science. *J. ACM* 50, 1 (Jan. 2003), 96–99. https://doi.org/10.1145/602382.602410

[36] Wen-Jing Yan, Q. Wu, Yong-Jin Liu, Su-Jing Wang, and X. Fu. 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–7.

[37] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. 2020. Convolutional Hierarchical Attention Network for Query-Focused Video Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (April 2020), 12426–12433. https://doi.org/10.1609/aaai.v34i07.6929 Number: 07.

[38] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. *PLOS ONE* 9 (01 2014), 1–8. https://doi.org/10.1371/journal.pone.0086041

[39] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442* (2019).