Multimodal Learning Analysis via Machine Learning and Deep Learning Methodologies

Taposh Kumar Neogy

Assistant Professor of Accounting, Department of Business Administration, Institute of Business Administration (IBA), Rajshahi (under National University), **BANGLADESH**

ABSTRACT

The world we live in is purely multimodal in nature. We can see objects, hear sounds, smell different sorts of scents, feel things, and taste various flavors. The word 'Modality' refers to something that occurs or something that can be experienced. An experience is presented as multimodal only when it includes some of the modalities found in the world. In order to gain recognition in understanding our general surroundings, artificial intelligence should be able to understand such multimodal perspectives. On the other hand, multimodal machine learning refers to the development of models that can interact and correlate data from various modalities. It's a dynamic and multi-disciplinary area of expanding significance with exceptional prospective. Rather than focusing on some limited multimodal applications, we, in this paper review the new technological developments in multimodal machine learning and present them in a typical scientific categorization. We move past the usual classifications and discuss more extensive opportunities and challenges presented by multimodal machine learning. Most of the studies conducted on multimodal learning methodologies utilize polls and surveys as their primary source of collecting quantitative data. This paper discusses the results of a precise literature of observational studies on the skills of multimodal data (MMD) for human learning with the help of artificial intelligence-empowered methodologies i.e. Machine Learning and Deep Learning. This paper also gives an outline of what and in what manner MMD has been leveraged to strengthen learning and in what environmental settings. The discussion of this paper portrays the abilities of multimodal learning and the continuous advances and methods that rise out of the work of multimodal learning to enhance and further elevate learning process. At last, we conclude that the future researchers should thoroughly consider developing a system that would empower multimodal features to be lined up with the ongoing research and learning plan. These features could likewise be used on enabling theory and practice activities to further elevate the multimodal learning process. This paper sets a clear pattern to enable the adoption of multimodal data inside future learning technologies and development endeavors.

Source of Support: Multimodal Data, Multimodal Machine Learning, Data, Multimodal Analytics, Machine Learning, Deep Learning

INTRODUCTION

Multimodal Learning alludes to the collection, measurement, analysis, and representation of information about students and the environment learning occurs, for understanding and improving the quality of learning and the infrastructure in which it happens (Ganapathy, 2016). The information related to traditional learning is typically uni-dimensional (Vadlamudi, 2017). For instance, only log data produced by a learning management system is only useful for just evaluating the learning process. In particular, this information overlooks the significance of context-based data about students (Bynagari, 2014).

This information related to context-based environment is critical for understanding students' learning methodologies. As such, uni-dimensional data give just incomplete data with regards to the learning system (Eradze & Laanpere, 2017) that makes it very challenging to deliver precise results of learning analysis (Ganapathy, 2017). The learning process is actually unpredictable (Bynagari, 2015). To analyze and depict a learning process precisely (Vadlamudi, 2016), we should gather multimodal information such as data related to the learning behavior, students' facial expression, and physiological information (Bynagari, 2015). Along these lines, a superior, a more universal approach of learning can be uncovered.

The world surrounding us includes various modalities. In simpler terms, the word modality refers to the occurrence of something or the way something is being experienced. The vast majority of people relate the word modality to sensory modalities. To make things simpler for you, sensory modalities are the type of modalities which address our essential channels of communication and feeling such as vision or touch. However, a research problem or database is subsequently labeled as multimodal when it includes numerous other modalities. Multimodal machine learning is designed to support and create models that can relate and extract data from various modalities. To accomplish this, we have discussed six different objectives of this multimodal research to help with understanding human learning.

Nonetheless, the research of multimodal machine learning comes along with some extraordinary opportunities and challenges for the field experts due to the heterogeneity of the data. Collecting data from multimodal sources and learning different approaches offers the chance of capturing communication among various modalities and acquiring a top-to-bottom representation of regular portents. In this paper we identify and discusses five mainly-specialized challenges presented by multimodal machine learning. Since these challenges are fundamental to the multimodal environment and should be handled to advance the field furthermore, we also discuss the opportunities multimodal learning has to offer once these challenges are overcome.

LITERATURE REVIEW

The conversion of multimodal information (MMD) with leading computational analyses empowers us to understand and support complex learning marvels (Blikstein and Worsley, 2016). For instance, eye-tracking information and the diverse semantic and prosodic elements of language can lead us to the students' skill (Andrade, Delandshere, and Danish, 2016; Mangaroska, Vesin, and Giannakos, 2019); or video information can enlighten us regarding their engagement (Nguyen, Huptych, and Rienties, 2018; Pardo, Han, and Ellis, 2016).

These experiences can empower noteworthy feedback to be given to the students. For instance, Hutt et al. (2019) utilized eye-tracking to automatically identify mind-wandering in online classes; while Grawemeyer et al. (2017) utilized students' speech and interaction to

figure out students' affective state. Such develops (eg, mind-meandering, emotional states) are utilized to give feedback to the students by assisting with determining the kind of input that should to be given – intelligent, educational – and how it should be presented – evaluative, interpretive, steady, probing).

The insights extracted from MMD allow us to explore students' behavior in ways that would not be possible with singular information sources. Giannakos, Sharma, Pappas, Kostakos, and Velloso (2019) concluded that the prediction of skill obtaining was much better with the help of MMD (eg, eye-tracking, Electroencephalography [EEG] and facial expressions) than when utilizing any singular stream. Research has shown that the combination of MMD brings essentially better prediction of learning results and assists us with interpreting complex learning curves (Giannakos et al ., 2019; Liu et al ., 2019; Sharma, Papamitsiou, and Giannakos, 2019; Sharma, Pappas, Papavlasopoulou, Giannakos, 2019).

However, the incredible ability of multimodal data in human learning, and late developments with regards to people and research conducted in this area has not arrived at its latent capacity. There are multimodal machine learning domains that are yet generally unattended. Along these lines, different researchers found that the performance of MMLA in the areas where learning happens is extremely difficult and generally restrictive with the outcome that MMLA's maximum capacity is significantly wasted.

LEARNING ANALYTICS METHODOLOGY IN MULTIMODAL ENVIRONMENT

A. Dependent Variables

Scientists have taken a closer look at various developments inside their particular research. These developments regularly lead to the hypothetical direction that the field experts are usually following. Regardless, the inclusion of dependent variables identified a few normal classes of dependent variables. The variables that appeared across various papers include: learning, behavior, collaboration, expertise, affect, attention, presentation, success. (Figure 1). Generally, the approach through which the analysts evaluate each one of these constructs is exceptionally variable. Some of these are based on the human coding, while other ones utilize heuristics approach. Likewise, analysts use various modalities to learn a similar develop. For instance, a few experts utilized speech to consider affect, while others utilized expressions.

B. Collaboration

While evaluating the previously mentioned dependent variables, around 22 percent considered the groups as the essential

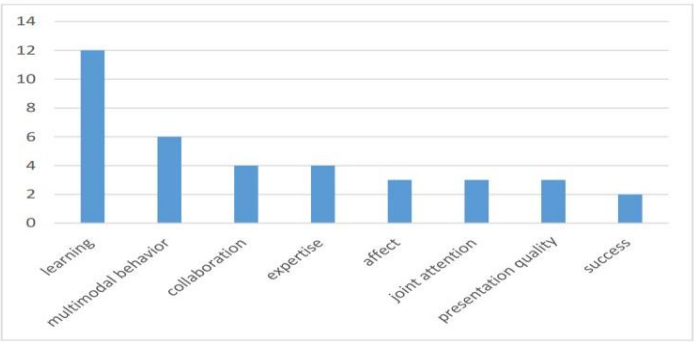

Figure 1: Various Dependent Variables Used in Different Research Papers

Element of analytics, 48 percent considered only the individual as a unit of analysis, and 30 percent considered both the individuals and groups as a unit of measurement.

C. Tools and Techniques

Some of initial data in multimodal landscape was developed in accordance with the insightful methods utilized to explore the various data streams. A few research studies were based on some specific codes, however majority of those used existing codes as well as tools to lead the research. Instances of such occurrences include: Linguistic Inquiry Word Count, OpenEAR, and FACET.

In different cases, scientists developed customized tools based on APIs and SDKs. Also, a lot of them utilized conventional ML and DL algorithms such as SVM (Support Vector Machine), Decision Trees, Bayesian Networks, etc. It is also worth noticing that a large numbers of these papers utilized hand-comments to develop supervised learning algorithms.

Table 1: Setting in a Multimodal Learning Environment Enabled by Machine Learning

	Feature Learning	**Supervised Learning**	**Testing**
Classic Deep Learning	Audio	Audio	Audio
Multimodal Fusion	Audio + Video	Audio + Video	Audio + Video
Cross Modality Learning	Audio + Video	Audio	Audio
Shared Representation Learning	Audio + Video	Video	Audio

CURRENT SCENARIO IN MULTIMODAL LEARNING LANDSCAPE

Addition of the existing literature in multimodal learning landscape depends on the phenomenon presented in some non-specific papers. Each of these documents was crafted according to the central ideas that were brought up by field experts throughout the research. This section presents an outline of some categories that are key pioneers of MMLA in current scenario.

A. Mobility

An interesting idea that was presented in a few papers is the approach we can use to capture user location utilizing their cell phones. Such information enables experts to contemplate members' real-time location and additionally gives a moderately simple way to collect geo-location, video and other multimodal information transfers. Moreover, this information has utility for analyzing educator development inside a learning environment, as well as understanding the relationship of various kinds i.e. student-student, student-innovation and student-teacher relationships.

B. Frameworks and Models

This idea reflected that traditional multimodal learning tools and models are the supplement of systems and models that provide better generalization and relevance. Simultaneously, using these models can assist with setting up standards for how information is analyzed across various environments, and help experts all the more simply arrange the objectives and direction of their research. At last, establishing models and frameworks can improve the formation of such frameworks and add expanded authenticity to multimodal models and tools.

C. Data Visualization

Field experts are additionally looking to tackle challenges related to data visualization, developing new data analysis devices, and integrating the existing information analysis tools with developed ones. While some initial research hypothesis and tools have been developed that assist with such errands, there is still a huge requirement for new and powerful tools for data visualizations. This section additionally includes concerns identified with data normalization, and the general simplicity of analyzing multimodal information. Scientists recommend using existing APIs, however, the information standards, data collection and data visualization are firmly associated with each other.

D. Human-Computer Analysis Collaboration

Another topic of interest is to direct research that considers the crossing point of human-computer joint effort. In particular, experts are searching for approaches to take advantage of human derivation and AI through bootstrapping human analysis with AI, or using human interaction within the data analysis pipeline.

E. Classroom Orchestration

This sort of classification considers current work to make the outcome of multimodal learning frameworks all the more significant. The information collected via this activity can be identified through student-teacher interfaces that members understand, and also through insightful frameworks to assist with arranging the right user learning experience.

F. Cross MMLA

The recent development in the Cross MMLA identifies one of the latest things inside the multimodal learning community. In particular, experts are progressively occupied with utilizing multimodal learning frameworks to improve the learning processes across various advanced and digital spaces. Directing a research activity of this scale brings various challenges as far as data collection, interoperability and normalization are being discussed. These classifications in no way, size or form address all of the research conducted in the field of multimodal learning analytics. Nonetheless, they surely address some of the most revolutionary ideas that are bound to be progressed by numerous domain experts inside this field.

MULTIMODAL DATA, LEARNING INDICATORS AND THE RELATIONSHIP BETWEEN THE TWO

A. Multimodal Data

A majority of the recent studies perceived the significance of multimodal information. But, few papers methodically arranged multimodal data types. In particular, this order structure comprised of digital domain, physical domain, physiological domain, psychometric domain, and environmental domain. Digital domain referred to different advanced methods created on the digital frameworks while training the learning system, like an online framework, virtual analysis framework, or STEAM framework.

Physical domain covers the information acquired by different gadget that leverage motion sensors and body movement. With the drastic improvement in the adoption of these AI-enabled sensors, the information acquired was more precise, refined and accurate at the miniature level, for example, the way students move their head (Bynagari, 2016), and the way they tap on the screen with their fingers (Ganapathy & Neogy, 2017). The insights and

analysis obtained with the help of this physical information were quite huge for the system to understand.

Physiological domain reflected the information identified with internal physiological reaction in a human, which showed students' learning status. Conversely, psychometric domain, a somewhat common stream of learning data, directed towards various different self-directed surveys that abstractly showed student's psychological state. Environmental domain represented the information related to a learning environment where students were actually present, such as temperature and weather. Research has demonstrated that an environment has huge effect on learning (Vadlamudi, 2015). The expanding significance of such environmental data is a need of the hour in multimodal learning.

Because of technical developments such as IoT (Internet of Things), cloud data storage, and wearable devices, data at a high-frequency and miniature level can be managed easily and precisely. From various measurements, multimodal learning frameworks powered by machine learning display students' actual learning state more efficiently and accurately (Bynagari, 2017), particularly in some specific courses (Ganapathy, 2015). Since learners connect with learning content, friends, and educators in a variety of ways, it is fundamental to analyze the learning processes through these multimodal data streams.

B. Learning Indicators

The main learning indicators utilized in the multimodal learning analytics are conduct, consideration, reasoning, metacognition, sentiment, teamwork, communication, commitment, and learning performance. We can classify some of them additionally. Specifically, learning conduct is further classified into three classifications — online learning, learning in the classroom, and epitomized learning conduct. Consideration includes individual consideration and joint consideration. Sentiments allude to those in self-sufficient learning and community learning. The teamwork comprises of real-time interaction process and remote teamwork through online mediums. Engagements refers to the commitment in learning and the face-to-face learning within classrooms. Summing it all up, the assessment score, the score of learning, is the common learning indicator for overall performance.

A few papers recommend various performance measurement techniques to work on the precision of the performance of learning indicators. Some utilize traditional techniques to assess this performance, such as, critical thinking skills. A few researchers focus on different parts of learning performance, like coordinated effort, task execution, and learning. Such skills incorporate verbal presentation and clinical activity skills. Through analyzing these learning indicators, experts concluded that a certain types of learning indicators are to be considered that denote to the complexity of the learning and the importance of some of the learning indicators mentioned here. For instance, a few experts led a different analysis of conduct, intellectual engagement, and sentiment in the learning system. In contrast, a few examinations consolidated these components to evaluate the learning process altogether.

Depending on the idea of commitment, specialists took notice of engagement with the help of various modalities of verbal nature as an indicator, kinesics as substantial commitment, and vocals as enthusiastic commitment. On the other hand, teamwork can be evaluated independently and one can measure sentiments in overall learning processes collectively.

There are a few guidelines for choosing learning indicators. Collaborative learning revolves around cooperative elements and team effort, while independent learning considers attention and commitment. Moreover, there are learning indicators of real-time interaction

with less remote engagement. With a thorough assessment of this multimodal learning system, learning indicators can be a bit more different.

C. The Relationship between Multimodal Data and Learning Indicators

Multimodal Learning Ecosystem makes a multi-dimensional research area to make the connection between data and learning indicators a lot more complex. Studies have discovered that there are three kinds of relationships between multimodal information and indicators: 1. One-to-one 2. Many-to-one 3. One-to-many

First relationship (1:1) implied that a sort of information was applicable to just estimate only one indicator. This is the most well-known type in multimodal machine learning space. As the technology has progressed impressively, measurement estimation capability of each sort of information is slowly tapped. Hence the 1:1 type of relationship has become extremely uncommon. Such as, the most widely recognized strategies to measure reasoning are online polls and interviews.

Through the new logic strategy, reasoning is evaluated by utilizing sound information. Since the physiological estimation is accessible now, data such as EEG information is also used to evaluate reasoning. Experts regard these techniques as the second type of the relationship. Second relationship demonstrated how many different types of data evaluate a similar learning indicator. For instance, ECG and EEG measure the level of involvement/engagement of students.

At last, the third type of relationships; that is, a singular type of information can measure a few multiple types of indicators. For instance, movement of eye measures how attentive a learner is, and so on.

The fundamental motivation behind why we have so many different corresponding relationships is that the scope of substantial indicators and nature of data change with specialized and hypothetical situations. As a rule, the evaluation scope of a specific type of information is pretty much restricted along with clear benefits.

However, there are more than one indicators with impactful measurement elements. Such as, online learning data is frequently leveraged to present learning conduct, while the movement of eye is often used to evaluate a student's intellectual level, and data handling process according to different learning content. On the other hand, expressions have a superior impact on sentiments and commitment. Expressions are a decent proportion of strong sentiments within a learning environment. Various research experts have claimed that an indicator can be measured via one-dimensional or multi-dimensional data streams. However, the estimation of such indicators should consider existing information as well as the combination of different types of information that is of importance to data integration.

MAJOR CHALLENGES IN MULTIMODAL MACHINE LEARNING SETTING

This section will discuss some basic concepts of the five fundamental challenges faced in multimodal machine learning environment:

A. Representation

A first basic challenges is to figure out how to address and represent the multimodal information to feature the complementarity and synchrony between different modalities. The variety of information makes it increasingly difficult for connected and joint

representations. For instance, language is frequently seen as figurative while sound and graphic ones will be addressed as indicators.

B. Translation

The next challenges is the way to decipher information starting with one methodology then onto the next. The obtained information is not only diverse, but the relationships are also usually abstract. Such instance, while representing a particular picture verbally, more than a single depiction can be right. The evaluation and translation of the multimodal information might be abstract.

C. Alignment

The next challenges is to figure out the connections between components from at least 2 unique modalities. Such as, while analyzing the speech and gestures of a human subject, how can we align specific signals to the verbally expressed words or expressions? This alignment between modalities might be based on long-range conditions and the classification is frequently ambiguous (e.g., words or expressions).

D. Fusion

The fourth challenge is to combine data from at least two modalities to perform a predictive task – discrete or continuous. For instance, the graphical representation of the lip movement is combined with the language signs to identify expressed words. The data extracted from various modalities might have differing power and noise levels. Multimodal fusion needs to deal with such varieties.

E. Co-learning

The fifth and final challenge is to transmit information among different modalities and their representations. Demonstrated by the methods used in machine learning, hypothetical grounding and zero-shot learning, how does training data from one methodology will be able to help a machine learning framework be trained on an alternate methodology? Hence, this is an especially important challenge when one of the modalities has restricted learning resources (e.g., labeled datasets).

	Challenges				
Applications	Representation	Translation	Alignment	Fusion	Co-learning
Speech recognition and synthesis					
Audio-visual speech recognition	✓		✓	✓	✓
(Visual) speech synthesis	✓	✓			
Event detection					
Action classification	✓			✓	✓
Multimedia event detection	✓			✓	✓
Emotion and affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media description					
Image description	✓	✓	✓		✓
Video description	✓	✓	✓	✓	✓
Visual question-answering	✓		✓	✓	✓
Media summarization	✓	✓		✓	
Multimedia retrieval					
Cross modal retrieval	✓	✓	✓		✓
Cross modal hashing	✓				✓

Figure 2: A Summary of Challenges for Each Application Enabled by Multimodal Machine Learning

OPPORTUNITIES PRESENTED BY MULTIMODAL FRAMEWORKS IN REAL WORLD

Adoption rate around driving multimodal applications into gadgets keeps on developing, with the following end-market verticals most probably ready.

In the automotive domain, MMLA is being related with ADAS (Advanced Driver Assistance Systems), HMI (In-Vehicle Human Machine Interface), and DMS (Driver Monitoring Systems) for immediate inference and forecast.

Advanced robotics merchants are actively integrating MMLA frameworks into HMIs and automation to expand product quality and consumer appeal that assist with more prominent partnership among workers and AI-enabled robotics in the modern landscape.

Consumer-oriented organizations in the cell phone and smart home business sectors, are in wild rivalry in order to show the value of their offering over their rivals'. MMLA elements and enhanced frameworks are basic to creating an advertising impact, making consumer hardware organizations prime contender for integrating multimodal learning-empowered frameworks into their gadgets. Developing applications include the verification of home security and installment verification and many more.

Healthcare organizations and clinics are yet somewhat new in the research and adoption of MMLA strategies, however there are certain encouraging arising use cases in medical imaging as of now. The importance of MMLA for doctors and patients will be a troublesome recommendation for medical administrations to oppose, regardless of whether adoption rate is relatively lower.

Entertainment organizations are now utilizing MMLA to assist with organizing their audio as well as video content into labeled metadata to further develop content recommendation frameworks, customized advertising, and automated compliance systems. Up until this point, installation of labeling frameworks have been restricted since MMLA innovation has lately been available for the sector.

CONCLUSION

As an ever increasing number of information related to various learning methodologies and processes become accessible, multimodal machine learning analytics has been turning out to be progressively significant. Multimodal learning can possibly connect the landscape of AI gadgets and genuinely empower business insight and endeavor huge advancement. This paper introduced a fundamental writing survey about Multimodal Learning Analytics. It utilized various basis to recognize trends and applications in multimodal machine learning, as well as promising circumstances for upcoming developments. While the methodology for identifying these papers can be stretched further to include more insights regarding specialized tools, the latest studies propose that over a wide span of time work in multimodal learning area has established a solid framework for on-going studies.

Significantly, analysts are claiming the methods to collect MMLA analytics data from students in different learning environments, and in order to lead research activities at individual and community levels. Progressing ahead, the research area seems to be ready to keep up in environmental stages, and to be extended towards extracting information across various domains. Moreover, people are probably going to experience the advancement of all the more robust systems, improved on data integration and tools. The experts can propel this research field by making the most of and adding to machine and deep learning. All the

more significantly, the field prosper by considering the approaches in which that multimodal machine learning will be able to add decisive value to convenience.

REFERENCES

Bynagari, N. B. (2014). Integrated Reasoning Engine for Code Clone Detection. *ABC Journal of Advanced Research*, *3*(2), 143-152. https://doi.org/10.18034/abcjar.v3i2.575

Bynagari, N. B. (2015). Machine Learning and Artificial Intelligence in Online Fake Transaction Alerting. *Engineering International*, *3*(2), 115-126. https://doi.org/10.18034/ei.v3i2.566

Bynagari, N. B. (2016). Industrial Application of Internet of Things. *Asia Pacific Journal of Energy and Environment*, *3*(2), 75-82. https://doi.org/10.18034/apjee.v3i2.576

Bynagari, N. B. (2017). Prediction of Human Population Responses to Toxic Compounds by a Collaborative Competition. *Asian Journal of Humanity, Art and Literature*, *4*(2), 147-156. https://doi.org/10.18034/ajhal.v4i2.577

Eradze, M. & Laanpere, M. (2017). Lesson Observation Data in Learning Analytics Datasets: Observata. In Proceedings of the 12th European Conference on Technology-Enhanced Learning (EC-TEL 2017), Tallinn, Estonia, pp. 504–508.

Ganapathy, A. (2015). AI Fitness Checks, Maintenance and Monitoring on Systems Managing Content & Data: A Study on CMS World. *Malaysian Journal of Medical and Biological Research*, *2*(2), 113-118. https://doi.org/10.18034/mjmbr.v2i2.553

Ganapathy, A. (2016). Speech Emotion Recognition Using Deep Learning Techniques. *ABC Journal of Advanced Research*, *5*(2), 113-122. https://doi.org/10.18034/abcjar.v5i2.550

Ganapathy, A. (2017). Friendly URLs in the CMS and Power of Global Ranking with Crawlers with Added Security. *Engineering International*, *5*(2), 87-96. https://doi.org/10.18034/ei.v5i2.541

Ganapathy, A., & Neogy, T. K. (2017). Artificial Intelligence Price Emulator: A Study on Cryptocurrency. *Global Disclosure of Economics and Business*, *6*(2), 115-122. https://doi.org/10.18034/gdeb.v6i2.558

Vadlamudi, S. (2015). Enabling Trustworthiness in Artificial Intelligence - A Detailed Discussion. *Engineering International*, *3*(2), 105-114. https://doi.org/10.18034/ei.v3i2.519

Vadlamudi, S. (2016). What Impact does Internet of Things have on Project Management in Project based Firms?. *Asian Business Review*, *6*(3), 179-186. https://doi.org/10.18034/abr.v6i3.520

Vadlamudi, S. (2017). Stock Market Prediction using Machine Learning: A Systematic Literature Review. *American Journal of Trade and Policy*, *4*(3), 123-128. https://doi.org/10.18034/ajtp.v4i3.521

--0--