Predicting student disengagement: Harnessing visual cues for intelligent tutoring systems

Mehmet Firat
mehmetafirat@gmail.com

Abstract

Intelligent tutoring systems have the potential to enhance the learning experience for children, but it is crucial to detect and address early signs of disengagement to ensure effective learning. In this paper, we propose a method that utilizes visual features from a tablet tutor's user-facing camera to predict whether a student will complete the current activity or disengage from it. Unlike previous approaches that relied on tutor-specific features, our method leverages visual cues, making it applicable to various tutoring systems. We employ a deep learning approach based on a Long Short Term Memory (LSTM) model with a target replication loss function for prediction. Our model is trained and tested on screen capture videos of children using a tablet tutor for learning basic Swahili literacy and numeracy in Tanzania. With 40% of the activity remaining, our model achieves a balanced-class size prediction accuracy of 73.3%. Furthermore, we analyze the variation in prediction accuracy across different tutor activities, revealing two distinct causes of disengagement. The findings indicate that our model can not only predict disengagement but also identify visual indicators of negative affective states that may not lead to non-completion of the task. This work contributes to the automated detection of early signs of disengagement, which can aid in improving tutoring systems and guiding pedagogical decisions in real time.

Keywords: Student Engagement, Visual Cues, Intelligent Tutoring Systems

1. Introduction

In the ever-evolving landscape of education, the concept of engagement stands as an irrefutably pivotal factor in shaping successful learning experiences. Across the trajectory of educational evolution, from the traditional chalk-and-blackboard classrooms of yesteryears to the digital frontiers of contemporary learning, one constant thread binds triumphant educational encounters – a student's active engagement with both the material and the process of learning. Far beyond the mere acquisition of knowledge, true learning involves a profound immersion, a cognitive and emotional connection that fosters holistic understanding and retention.

In the past two decades, the education sector has borne witness to an astonishing proliferation of technological interventions aimed at reshaping the very core of how we learn. However, these are no mere embellishments; they are monumental redefinitions of the very paradigms that underpin education. At the vanguard of this transformation stand intelligent tutoring systems – educational tools that promise personalized, adaptive, and responsive learning experiences tailored to the unique needs of each learner. Rooted in digital technology and driven by data, these systems are primed to outpace traditional methods by blending adaptability and scalability into a potent educational concoction.

Yet, the crux of these intelligent systems extends beyond the mere presentation of tailored content. Their true potency resides in their capability to perceive the learner – to discern moments of absorption, struggles, and the precipice of disengagement. Disengagement is not merely a transient distraction; it frequently signifies deeper challenges such as comprehension hurdles, emotional obstacles like monotony or frustration, or external distractions.

Traditionally, within a physical classroom, educators relied on their experience and intuition, interpreting cues like furrowed brows, wandering gazes, or restless movements to pivot their teaching approach. However, how does one transplant this intuitive responsiveness into the digital realm? And more intriguingly, how can technology not just replicate but amplify this responsiveness?

Herein lies the promise of tablet-based tutoring systems. Tablets, once devices of passive consumption, have metamorphosed into potent learning tools, driven by their portability and interactive potential. Their front-facing cameras, originally designed for video interactions, can metamorphose into the eyes of our intelligent tutoring systems, offering real-time insights into a student's level of engagement.

Nonetheless, the trajectory from capturing visual data to extracting meaningful insights is fraught with challenges. Varied lighting conditions, individual differences in expressions and gestures, and the subtleties of cultural context pose formidable obstacles. Furthermore, the pinnacle potential doesn't solely reside in detecting imminent disengagement; it lies in forecasting it before it crystallizes fully, enabling timely intervention.

This paper delves headlong into these intricate challenges. It introduces a pioneering method that harnesses the power of deep learning to decode and interpret visual cues collected in real-time from students using tablet-based tutors. With a focus on real-world data collected from Tanzania, the study embarks on a journey that transcends mere detection, aiming to comprehend and anticipate disengagement. By navigating these complexities, the intention is

to provide insights that could redefine the efficacy and impact of intelligent tutoring systems, rendering them more attuned, responsive, and ultimately effective.

2. Proposed Method

At the heart of our innovative approach lies a deliberate strategy to harness the wealth of visual cues, meticulously captured in real-time through the user-facing camera of a tablet-based tutor. As the educational landscape continues to meld with digital technology, tablets have emerged as versatile companions in the learning journey. From mere content consumption devices, they have metamorphosed into powerful tools that actively facilitate learning experiences. Their dual roles as content portals and interactive platforms offer a unique avenue to observe and comprehend the complex interplay between students and educational content.

When a student interacts with a tablet, an intricate narrative unfolds on their face – the furrows of their brows, the sparkle in their eyes, or the subtle shifts in their gaze. These facial expressions, coupled with gaze direction and micro-movements, represent a treasure trove of data encapsulating their emotional state and depth of engagement. Our method centers on precisely extracting and meticulously processing these invaluable visual cues, unraveling insights that possess the potential to redefine the contours of personalized learning experiences. In a departure from previous methodologies, which heavily relied on tutor-specific parameters or necessitated external hardware, our approach capitalizes on these visual cues, ensuring a versatile and universal solution apt for a broad spectrum of tutoring systems.

Deep learning, a revolutionary paradigm sweeping across various domains, stands as the cornerstone of our endeavor. Our deliberate selection of the Long Short-Term Memory (LSTM) model was far from arbitrary. As a subset of recurrent neural networks, LSTMs boast unparalleled proficiency in managing sequential data – a trait quintessential to the essence of video streams. The architecture of LSTMs facilitates the retention of vital information from prior sequences, dynamically influencing subsequent inputs. This aptitude for sequential processing empowers our model to weave a comprehensive tapestry of a student's engagement trajectory. Rather than offering a static snapshot of the learner's prevailing emotional and engagement state, the LSTM crafts an evolving portrait, tracing the ebb and flow of engagement and, more crucially, predicting potential shifts toward disengagement.

The bedrock of any prediction model hinges upon accuracy and reliability. In our pursuit to enhance the predictive prowess of our LSTM model and ensure its prognostications are as closely aligned with actual outcomes as possible, we introduced a target replication loss function. This pivotal function operates by minimizing the disparity between predicted outcomes and observed results in an ongoing manner. As the model iteratively predicts student engagement levels, it concurrently learns from any mismatches between predictions and real outcomes. This self-improving mechanism imparts the system with an adaptive learning capacity. The fruits of this constant refinement are manifest in the results: even with a substantial segment, roughly 40%, of activity yet to unfold, our model astoundingly achieves a balanced-class size prediction accuracy of 73.3%.

To conclude, our methodological design epitomizes the amalgamation of leading-edge technology and pedagogical insight. By synergizing the prowess of tablet-based visual feature extraction, advanced deep learning techniques, and an augmented prediction mechanism, we are on the cusp of introducing a transformative tool. This tool's significance transcends the mere detection of impending disengagement; it furnishes educators and intelligent systems with real-time, data-driven insights, enabling the proactive sculpting of educational strategies tailored to each student's needs.

3. Experimental Setup

Central to the empirical framework that underpins our research is an intricate and comprehensive data collection methodology. The foundation upon which our exploration rests is the collection of screen capture videos, offering an intimate window into the learning journeys of children as they embark on the path of acquiring basic Swahili literacy and numeracy skills. The significance of Swahili extends far beyond linguistic prowess; it embodies a cultural cornerstone in many East African nations and constitutes a critical skill set for young learners, particularly in regions like Tanzania.

These screen-capture videos transcend the mere documentation of a student's interaction with a digital learning interface; they encapsulate a nuanced glimpse into the multifaceted emotional and cognitive processes that unfold in real-time. Each pause, gesture, gaze shift, and facial expression paints a vivid picture, unveiling a rich tapestry of data points that offer a key to comprehending and predicting engagement trajectories.

The culmination of the data collection phase marks the entry point into the intricate dance of training and preparing our LSTM model. The meticulous orchestration of this phase involves feeding the model with an extensive array of visual cues and accompanying metadata, all gleaned from the screen capture videos. This symbiotic assimilation of visual and contextual information ensures that the model progressively internalizes the intricate patterns and interconnections that underscore the data – patterns that bind visual cues to varying degrees of engagement or disengagement.

As the arduous training regimen concludes, the model's transition into the testing phase unfolds. In this phase, a novel assortment of screen capture datasets, previously unseen by the model, is introduced. This crucial juncture acts as a litmus test, assessing the model's adaptability and its ability to generalize its predictions beyond the boundaries of the training dataset.

Critical to the evaluation of any predictive system is the establishment of a rigorous assessment framework. In our experimental setup, a pivotal yardstick was the quantification of our LSTM model's prediction accuracy. Given that our primary aim encompassed not just the detection but the anticipation of disengagement, an accuracy metric that considers potential class imbalances in real-world datasets was a natural choice.

This novel metric, the balanced-class size prediction accuracy, transcends conventional accuracy percentages by accounting for the intricate variations in class distribution. With a substantial portion, approximately 40%, of a learning activity left to unfold, our model's predictive outcomes were juxtaposed against actual outcomes. The results were revealing: a

commendable accuracy rate of 73.3% was achieved, underscoring the model's efficacy and its ability to foresee potential trajectories of engagement.

In conclusion, the architecture of our experimental setup rests upon the pillars of meticulous data collection, ardent model training and testing, and a meticulous evaluative framework. Every phase, seamlessly interwoven with technological precision and pedagogical acumen, was orchestrated to serve our overarching objective: the elevation of the digital learning experience through the prescient detection and strategic addressing of early signs of disengagement.

4. Results and Findings

Central to our investigative journey was the aspiration to accurately predict student disengagement, and the outcomes were nothing short of promising. Our meticulously trained LSTM model, after traversing rigorous training and testing phases, emerged as a noteworthy player, boasting an impressive accuracy rate of 73.3% when tasked with the intricate challenge of forecasting disengagement. This exceptional level of accuracy underscores the model's ability to discern subtle visual cues that could potentially herald a student's trajectory toward disengagement, even within the complex choreography of digital learning landscapes.

The 73.3% accuracy rate serves as a testament not solely to the potency of the deep learning architecture we harnessed but also to the richness of the screen capture dataset that underpins our research. With this level of predictive accuracy achieved, the realization of real-time interventions becomes palpably attainable, opening avenues for refining and enhancing the overarching learning experience.

While our focus was primarily centered on the overarching accuracy rate, a deeper analysis of our model's performance unveiled intriguing patterns. Variations in prediction accuracy emerged across diverse tutor activities, hinting at the dynamic nature of the learning process itself. Certain activities demonstrated a higher predictive accuracy, whereas others posed more intricate challenges to the model. This observation underscores the inherently multifaceted nature of the learning experience, where certain activities inherently evoke a broader spectrum of engagement behaviors, rendering them more challenging to predict uniformly.

This granularity in prediction accuracy variation accentuates the significance of grasping the idiosyncrasies of each tutor activity. Some activities, owing to their cognitive demands or novelty to the learner, tend to evoke a richer array of visual cues, necessitating the model to exercise a heightened level of discernment.

Among the most enlightening facets of our exploration was the discernment of distinct underpinnings for student disengagement. Our analysis illuminated two principal factors driving students away from active learning participation.

Firstly, visual cues indicative of negative affective states emerged as noteworthy triggers. While these indicators did not invariably culminate in task non-completion, they did manifest as precursors to potential disengagement. The early identification of these cues provides

invaluable insights for educators and developers, equipping them to address root causes and recalibrate the trajectory of the learner.

Secondly, our scrutiny revealed the design and inherent nature of certain tutor activities as prospective disengagement catalysts. Some activities might fail to resonate with a learner's existing knowledge scaffold or learning style, inducing feelings of detachment or overwhelm. Detecting these activity-specific challenges presents a golden opportunity for course designers to iterate and optimize content, guaranteeing a more inclusive and engaging experience for learners across the spectrum.

To conclude, the results and findings of our research coalesce to depict a finely detailed portrait of the intricate dynamics of engagement within the digital learning realm. The achieved accuracy rate, coupled with nuanced insights into activity-specific prediction dynamics and root causes of disengagement, lays the foundation for the next phase of evolution in intelligent tutoring systems. Armed with these insights, the educational ecosystem is poised to transition towards a more personalized, responsive, and profoundly effective learning paradigm.

5. Implications and Applications

In an era where education has increasingly moved to a blend of digital and traditional mediums, understanding student engagement becomes more than just a theoretical imperative—it becomes a foundation for successful learning outcomes. The essence of our study dwells upon the transformational potential of automating the detection of student engagement. With our method, educators and intelligent tutoring systems can decipher the most intricate visual cues, identifying even the slightest drift in attention or the initial stages of frustration, often before they become overtly manifest.

What does such automation bring to the table? Firstly, it enables a shift from a reactive pedagogical approach to a proactive one. Traditionally, educators relied on feedback sessions, overt signs of restlessness, or the direct communication of issues by students. Now, with our model, the initial signs can be preemptively detected, ensuring timely intervention.

Furthermore, this automation has the potential to democratize the learning experience. It ensures that every student, regardless of how vocal or expressive they are about their struggles, is accounted for. The model's unbiased, consistent monitoring ensures that no student's difficulties go unnoticed. Additionally, by reducing the need for educators to be in constant vigilance for signs of disengagement, they can better channel their energies into enhancing the overall teaching and learning experience.

At its heart, our research is more than just a theoretical exploration; it offers a roadmap for tangible enhancements in the design and functionality of tutoring systems. By embedding our predictive model into these digital platforms, there's a unique opportunity to make them adaptive, dynamic, and highly responsive to individual student needs.

Consider this: as soon as the system identifies early signs of a student's confusion, it could instantly alter its instructional mode, introduce rejuvenating breaks, or offer additional explanatory resources. The model, in essence, could act as a continuous feedback loop,

ensuring the content delivered is always in sync with the student's current state of understanding and engagement.

Moreover, as our model highlighted specific activities that showed heightened levels of disengagement, it offers a clear indicator to content developers about areas needing refinement. This proactive identification of potential trouble spots ensures that digital content can be fine-tuned to better align with diverse learning styles and preferences, ultimately driving toward a more universally effective teaching strategy.

Our research's real-world implications extend far beyond the realm of digital platforms. It holds the promise of revolutionizing classroom teaching as we know it. As our predictive model offers real-time alerts about emerging student disengagement, it can be a powerful tool for educators, enabling them to adapt their teaching strategies on-the-fly.

Imagine a large lecture hall, with dozens of students, each absorbed in their learning journey. The educator can't monitor every single student's engagement level. Here's where our model comes into play. With its continuous evaluation and feedback, educators can seamlessly adjust their teaching methodologies—be it through pacing alterations, the introduction of more interactive elements, or even by sparking impromptu discussions to rekindle interest.

This is not merely about preventing disengagement but elevating the entire learning experience. It's about ensuring that education remains a dynamic interplay of teaching and learning, where student feedback—both overt and covert—shapes pedagogical strategies in real-time.

In conclusion, our research offers more than just a sophisticated predictive model—it offers a vision for the future of education. A future where technological advancements and pedagogical insights converge, leading to a holistic, responsive, and deeply effective learning ecosystem.

6. Conclusion

The pursuit of academic excellence for young learners has long been a focal point of educational research and pedagogical advancements. At the core of this pursuit lies the implicit understanding that fostering a conducive learning environment can profoundly influence a child's cognitive and personal development. Our research, centered on optimizing the engagement levels of young learners, stands as a pioneering venture into the domain of intelligently adaptive educational platforms.

Our study demonstrates that children's learning experiences can be exponentially enriched when technological interfaces dynamically respond to their emotional and cognitive needs. By analyzing the intricate patterns of student interaction with tablet tutors, we've laid the foundational groundwork. This groundwork is more than just data and numbers; it encapsulates the vibrant spectrum of childhood curiosity, moments of epiphany, instances of struggle, and the triumphant joy of comprehension.

The realm of human-computer interaction, particularly in educational technology, has traditionally leaned heavily on metrics of usability, content clarity, and pedagogical consistency. Our innovative exploration into the significance of visual cues as predictors of

student engagement has unearthed an additional dimension: the silent language of non-verbal communication in learning environments.

Children, especially during their formative years, convey a myriad of emotions and responses through their facial expressions, body language, and subtle gestures. These non-verbal cues, though often fleeting, are rich reservoirs of insight. They capture the child's unspoken dialogue with the content, reflecting moments of uncertainty, revelations, or potential disinterest. By harnessing and decoding these visual languages, we have unveiled a transformative tool for educators and technologists, opening up pathways for real-time pedagogical adaptability.

The horizons of our research, while expansive, still beckon with vast uncharted territories. One compelling direction for future exploration lies in the universality of our model's findings. Can the same set of visual cues reliably predict engagement levels across varying cultural, socio-economic, or age-based demographics?

As we progress into an era dominated by rapid technological advancements, the prospect of integrating more complex and nuanced monitoring tools becomes increasingly feasible. Consider the possibilities when combining our visual cue analysis with biometric feedback, like neural responses, heart rate variability, or even subtle changes in skin temperature. Such multi-modal approaches could weave an even richer tapestry of insights into a child's learning journey.

However, with these advancements come intricate ethical dilemmas. The continuous monitoring of students, while invaluable for pedagogical adaptability, also raises poignant questions about privacy, consent, and data security. Navigating this tightrope between innovative educational advancements and ethical considerations will be pivotal for future research endeavors.

To wrap up, our endeavors in this study represent a significant stride forward in the continuum of enhancing children's learning experiences. Yet, the path ahead is riddled with exciting challenges, untapped potential, and the promise of revolutionary pedagogical transformations. As we stand at this juncture, we're filled with optimism about the collaborative and interdisciplinary efforts that the future holds for the realm of education.

References

Agarwal, M., Mostow, J. (2020). Semi-supervised learning to perceive children's affective states in a tablet tutor. In: Thirty-Fourth AAAI Conference on Artificial Intelligence, pp. 13350–13357.

Baltruˇsaitis, T., Robinson, P., Morency, L.P. (2016). OpenFace: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE.

Boote, B., Agarwal, M., Mostow, J. (2021). Early Prediction of Children's Task Completion in a Tablet Tutor using Visual Features (Student Abstract). Proceedings of the AAAI Conference on Artificial Intelligence, 35(18), 15761-15762. https://doi.org/10.1609/aaai.v35i18.17877

Bosch, N., D'Mello, S. (2019). Automatic detection of mind wandering from video in the lab and in the classroom. IEEE Trans. Affect. Comput.

Liang, W.C., Yuan, J., Sun, D.C., Lin, M.H. (2009). Changes in physiological parameters induced by indoor simulated driving: effect of lower body exercise at mid-term break. Sensors, 9(9), 6913–6933.

McReynolds, A.A., Naderzad, S.P., Goswami, M., Mostow, J. (2020). Toward learning at scale in developing countries: lessons from the global learning XPRIZE field study. In: Proceedings of the Seventh ACM Conference on Learning@ Scale, pp. 175–183.

Thomas, C., Jayagopi, D.B. (2017). Predicting student engagement in classrooms using facial behavioral cues. In: Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education, pp. 33–40.