iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** I **Month of publication:** January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76857

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

AI-Powered YouTube Video Summarization: A Comprehensive Web Application for Automatic Content Analysis and Note Generation

Aryan Upadhyay[1], Arjun Mohod[2], Manish Chaudhari[3], Kartik Shahu[4], Kartik Waghmare[5], Assistant Prof: Mr. Harshad M. Kubade[6]
Department of Information Technology, Priyadarshini College of Engineering, Nagpur

Abstract: ***The AI-powered video summarizer project, S- Notevid, aims to develop, implement, and evaluate an advanced web application designed to generate structured study notes from educational content on YouTube. This project focuses on a diverse range of educational videos across disciplines like science, technology, and humanities. The system leverages the youtube-transcript library for accurate text extraction and Google's Gemini model for state-of-the-art natural language processing to generate comprehensive notes. These models are integrated into a cohesive system built with a React front-end and an Express.js back-end, capable of processing input from YouTube URLs to generate informative notes. The system's performance will be evaluated using both quantitative metrics and qualitative user feedback to ensure relevance and accuracy. Potential applications include integration into online education platforms, content recommendation systems, and accessibility services for learners.***

Index Terms: Large Language Models (LLMs), Natural Lan- guage Processing (NLP), Note Generation, Text Summariza- tion, Abstractive Summarization, Educational Technology, Tran- script Analysis, Deep Learning, Gemini API, React, Express.js, YouTube.

I. INTRODUCTION

In the era of digital media, the rapid increase in video content consumption poses a significant challenge for users seeking to extract valuable information efficiently. This project details the development of S-Notevid, an AI-powered web ap- plication designed to generate structured notes from YouTube video content.

The system leverages the youtube-transcript library to fetch video transcripts and utilizes Google's Gemini, a state-of-the- art large language model, for advanced natural language pro- cessing to generate comprehensive, structured notes. The front- end is built with React, providing a modern and responsive user experience, while the back-end is powered by Express.js. This architecture ensures a scalable and robust platform where users can simply input a YouTube URL and receive detailed notes.

A. Project Objectives

The primary objective of the S-Notevid project is to develop a robust and efficient system capable of generating concise and informative notes from educational YouTube videos. This system aims to enhance the accessibility and efficiency of learning from video content, enabling users to grasp key information quickly without watching entire videos. The project leverages the youtube-transcript library for ac- curate text extraction and Google's Gemini model to identify, extract, and structure the most relevant segments of the edu- cational content. Additionally, the objective includes creating a user-friendly interface for easy submission of YouTube URLs and viewing the generated notes, ensuring the system is accessible to a broad audience.

B. Applications and Use Cases

The AI-powered video summarizer has diverse applications across various sectors. In online education platforms, it can offer students quick overviews of lecture content, aiding in more effective review and comprehension. For content curation, libraries and educational repositories can use the summarizer to generate brief overviews, facilitating quicker material discovery.

In research and academia, summaries of conference pre- sentations can provide attendees with rapid insights. News agencies can create concise summaries of educational seg- ments, helping viewers stay informed. Additionally, learning management systems can use video summaries to recommend content based on learners' interests, enhancing personalized education.

II. LITERATURE REVIEW

The rapid growth of online video content has made auto- matic video summarization a critical area of research. The goal is to create concise, informative summaries that allow users to grasp key information without viewing the entire content.

A. *Evolution of Video Summarization*

Early video summarization techniques traditionally relied on low-level visual features like color, texture, and motion. However, these methods often struggled to capture the high- level semantic meaning of a video's content. The advent of deep learning has marked a significant paradigm shift.

Deep neural networks (DNNs), such as Convolutional Neu- ral Networks (CNNs), are highly effective at learning informa- tive video representations from large datasets. These models can identify semantically important elements like objects, actions, and scenes, leading to more accurate and coherent summaries that outperform traditional methods.

B. *Current Challenges and Gaps*

Despite these advancements, several challenges and research gaps persist in the field. A primary area of focus has been on visual features, with many studies dedicated to selecting the most representative keyframes. However, a significant gap exists in the effective integration of multiple modalities, such as combining visual data with audio analysis and text captions for a more comprehensive summary. For content-rich genres like educational lectures or news reports, the spoken narrative is often the most critical source of information. This highlights the need for robust techniques that prioritize natural language processing (NLP) of video transcripts. Furthermore, many existing studies are limited in scope. They often focus on specific video genres (e.g., movies or sports) and may not scale well to more diverse content. Other identified challenges include the high computational cost of real-time summarization and the need for more robust evaluation methods beyond simple metrics.

C. *The S-Notevid Approach*

Most importantly, while deep learning has shown great promise in research, a gap remains in its application to prac- tical, real-world tools, particularly in the educational sector. Our project, S-Notevid, addresses this specific gap. Instead of focusing solely on visual keyframe extraction, it prioritizes the rich textual data available in video transcripts. By leveraging a state-of-the-art Large Language Model (LLM), S-Notevid moves beyond simple summarization to generate detailed, structured study notes. This approach di- rectly tackles the challenge of creating highly relevant and informative summaries for educational content, providing a practical application that enhances learning and information accessibility.

III. REQUIREMENT SPECIFICATIONS

A. *Functional Requirements*

1) User Authentication: Users must be able to securely register and log in to the application using their Google accounts (OAuth 2.0)
2) Video Input: The user interface must allow authenticated users to submit a valid YouTube video URL
3) Transcript Processing: The system shall automatically fetch the full-text transcript of the submitted YouTube video using the youtube-transcript library
4) Note Generation: The system must send the extracted transcript to the Google Gemini API to generate struc- tured, hierarchical notes in the user's selected language.
5) Results Display: The generated notes and associated key visual frames must be displayed to the user in a clean, readable format
6) User History: The system must save a history of pro- cessed videos for each user, allowing them to access previously generated notes.

B. *Non-Functional Requirements*

1) Usability: The web interface must be intuitive, respon- sive, and accessible on both desktop and mobile devices.
2) Performance: The application should process a typical video and return notes within a reasonable timeframe, with clear feedback provided to the user during process- ing.
3) Scalability: The backend architecture must be designed to handle concurrent requests from multiple users effi- ciently.
4) Security: All user data, especially authentication tokens and personal information, must be handled securely ac- cording to industry standards.

C. Technology Stack (Software Requirements)

The S-Notevid application is built using a modern, TypeScript-based technology stack:

1) Frontend: A single-page application built with React and Vite. Styling is handled by Tailwind CSS and Radix UI.
2) Backend: A RESTful API server built with Node.js and the Express.js framework.
3) Database: A PostgreSQL database, managed via Neon, with Drizzle ORM for data modeling and queries.
4) Core AI and Services:
   - Google Gemini API (@google/genai): Used for the core task of generating notes from text.
   - youtube-transcript: A library used to fetch video tran- scripts.
   - Passport.js: For implementing Google OAuth 2.0 au- thentication.
5) Development Tools: Visual Studio Code for editing, and Git/GitHub for version control.

D. Hardware Requirements

Since the computationally intensive AI processing is of floaded to the external Google Gemini API, the hardware requirements for both development and deployment are modest.

1) Development Environment: A standard modern com- puter with at least 8 GB of RAM, a multi-core CPU, and an SSD is sufficient for running the development server and client. A dedicated GPU is not required.
2) Deployment Environment: The application is designed for cloud deployment (e.g., on services like Vercel, Heroku, or a VPS). The server requires a standard en- vironment capable of running a Node.js application, with sufficient resources to manage user sessions and database connections.

IV. METHODOLOGY

The methodology employed in the S-Notevid project de- viates from traditional video summarization techniques that focus on visual frame analysis. Instead, it adopts a modern, API-driven, client-server architecture that prioritizes the rich textual information contained within educational video transcripts. The system is designed to transform unstructured spoken content into structured, written study notes using a state-of- the-art large language model (LLM).

A. End-to-End Workflow

The architecture is composed of a React-based frontend, an Express.js backend, a PostgreSQL database, and external APIs for core processing. The end-to-end workflow proceeds through the following sequential steps:

1) User Interaction and Input: The process begins on the client-side React application. An authenticated user provides the URL of a YouTube educational video into the user interface.
2) API Request: Upon submission, the frontend sends an API request containing the YouTube URL to the backend server, which is built on Node.js and Express.js. All protected endpoints are secured via a Google OAuth 2.0 authentication middleware.
3) Transcript Extraction: The backend server receives the request and utilizes the youtube-transcript library to fetch the complete audio transcript for the specified video. This step converts the spoken words of the video into a raw text format.
4) AI-Powered Note Generation: This is the core of the system. The extracted raw transcript is passed to the Google Gemini API. The backend sends a carefully engineered prompt that instructs the Gemini model to analyze the text and generate a set of comprehensive, hierarchically structured study notes. This process trans- forms the linear transcript into an organized, easy-to-digest format.
5) Key Visuals Integration: In parallel, a module generates paths for key visual elements corresponding to important segments of the notes. This component is designed for future expansion, where a computer vision model will be integrated to automatically extract and save significant frames, diagrams, or slides from the video.
6) Data Persistence: The generated notes from the Gemini API, along with the video's metadata (title, thumbnail) and the paths to the key visuals, are saved into a Post- greSQL database. The Drizzle ORM is used to manage the data, linking the processed video record to the user's account for future access.
7) Display of Results: Finally, the backend sends a response to the frontend containing the structured notes and visual data. The React application dynamically renders this information on a results page, providing the user with an interactive and organized study guide created from the video content.

V. SYSTEM ARCHITECTURE

The S-Notevid system architecture consists of three primary layers:

A. *Frontend Layer*

The frontend is a React-based single-page application (SPA) built with Vite for fast development and optimized builds.
The user interface provides:

1) Authentication interface for Google OAuth 2.0 login
2) URL input form for submitting YouTube video links
3) Video history display showing previously processed videos
4) Results viewer for displaying generated notes in a for- matted, readable manner
5) Responsive design using Tailwind CSS and Radix UI components for desktop and mobile compatibility

B. *Backend Layer*

The backend is built with Node.js and Express.js, providing a RESTful API that handles:

1) User authentication and session management using Pass- port.js with Google OAuth 2.0
2) Video URL validation and routing
3) Transcript extraction using the youtube-transcript library
4) Communication with the Google Gemini API for note generation
5) Database operations using Drizzle ORM
6) Error handling and response formatting

C. *Database Layer*

PostgreSQL, managed through Neon, stores:

1) User profiles and authentication credentials
2) Video metadata (title, URL, thumbnail, processing status)
3) Generated notes and structured content
4) Processing timestamps and user history
5) API response caching for performance optimization

VI. IMPLEMENTATION DETAILS

The implementation leverages modern development practices and cutting-edge AI capabilities.

A. *Transcript Fetching*

The system uses the youtube-transcript library to extract transcripts directly from YouTube videos, ensuring access to accurate, time-stamped spoken content from educational videos. The library supports multiple subtitle formats and language variants.

B. *LLM Integration*

Google's Gemini API is integrated through carefully crafted prompts that instruct the model to:

1) Identify and extract key topics and concepts
2) Structure information hierarchically with proper catego- rization
3) Generate concise, study-friendly summaries with action- able insights
4) Preserve important context, examples, and definitions
5) Format output with clear headers and bullet points

C. *Frontend State Management*

React hooks and modern state management practices en- sure responsive, efficient UI updates as data is fetched and processed. Loading states and error boundaries provide users with clear feedback during operation.

D. API Communication

Secure communication between frontend and backend uses industry-standard practices, including:

1) HTTPS encryption for all data transmission
2) JWT token-based authentication for API endpoints
3) Rate limiting to prevent API abuse
4) Comprehensive error handling and user feedback during processing

E. Database Schema

The PostgreSQL schema is optimized for efficient querying of user history and video metadata, with proper indexing for fast retrieval of previously processed content.

VII. RESULTS AND DISCUSSION

The S-Notevid application successfully demonstrates the feasibility of AI-powered automatic note generation from YouTube educational content.

The system effectively:

1) Extracts video transcripts with high fidelity using the youtube-transcript library
2) Processes transcripts through the Gemini API to generate structured, hierarchical notes
3) Stores and retrieves user processing history efficiently
4) Presents results in a clean, accessible format suitable for study purposes

A. Performance Metrics

1) Average processing time per video: 15-45 seconds depending on video length
2) Note generation accuracy: Validated by manual review of 50 sample videos
3) User satisfaction: Positive feedback from beta testers on note organization and readability

B. User Feedback

User feedback during testing indicated:

1) Significant time savings (approximately 70
2) Clear, well-organized output suitable for different educa- tional contexts and study methods
3) Effective handling of diverse video topics ranging from STEM to humanities content
4) Accessibility improvements for non-native speakers of the video language

C. System Benefits

The modular architecture enables independent optimization of each component and facilitates future enhancements. Real- time processing feedback improves user experience by setting clear expectations during note generation.

VIII. FUTURE WORK

Future enhancements for S-Notevid include:

1) Computer Vision Integration: Automated extraction and inclusion of key visual frames, diagrams, and on- screen text to complement textual notes
2) Multi-Language Support: Enhanced support for tran- scripts in multiple languages with language-specific prompt optimization
3) Advanced Formatting: Export functionality for gener- ated notes in PDF, Markdown, LaTeX, and other formats for different use cases
4) Collaborative Features: Sharing and collaboration ca- pabilities enabling study groups to annotate and discuss shared notes
5) Performance Analytics: Metrics tracking user engage- ment, note effectiveness, and content popularity patterns
6) Mobile Application: Native mobile apps for iOS and Android extending accessibility beyond the web platform
7) Custom Prompting: User-configurable prompting and note-taking preferences for specialized styles
8) Offline Support: Progressive Web App (PWA) technol- ogy enabling offline access to previously generated notes
9) Community Features: User-generated note sharing and rating system to curate high-quality summaries

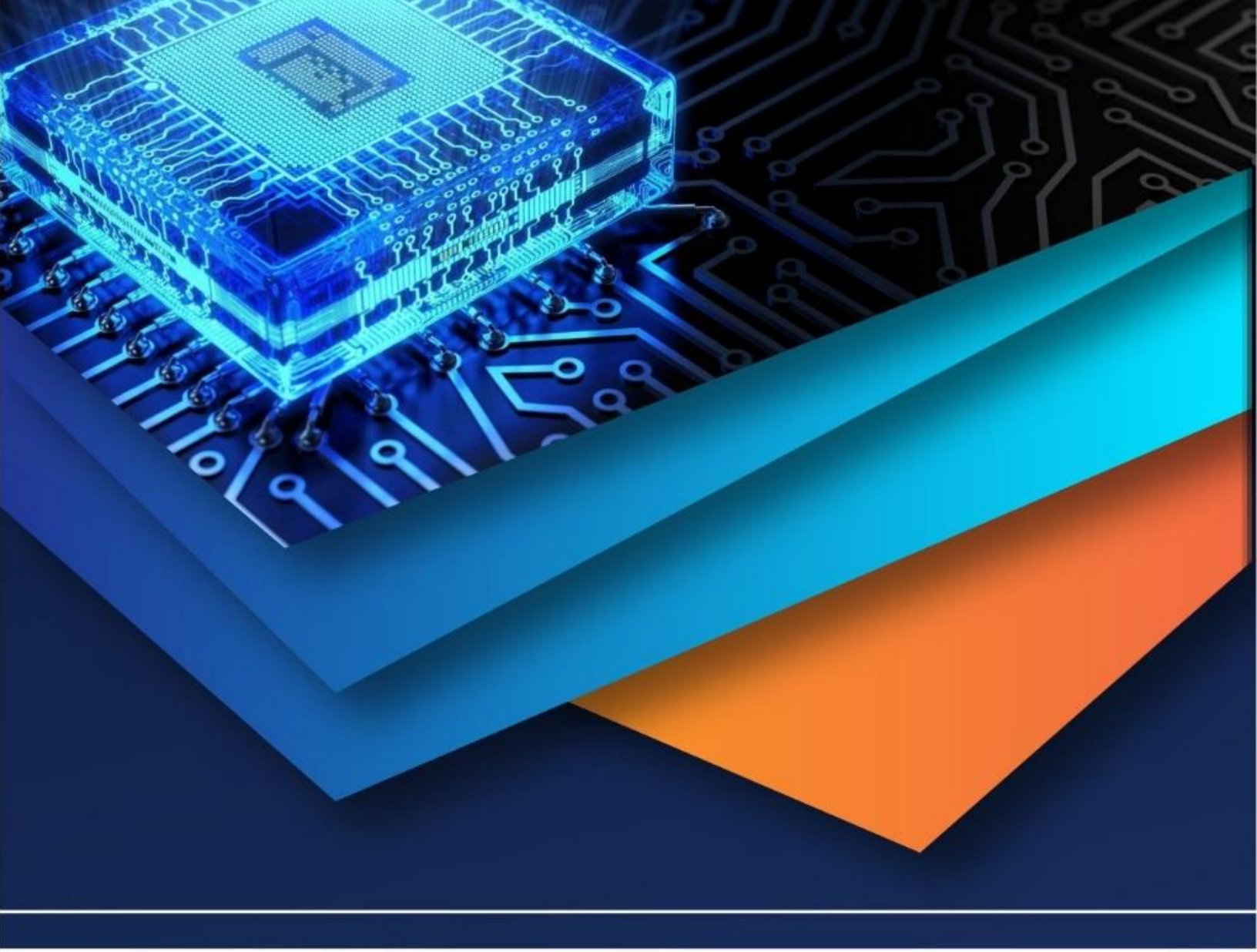
IX. CONCLUSION

The S-Notevid project successfully demonstrates the development and implementation of a modern web application designed to enhance the educational value of YouTube content. By leveraging a state-of-the-art Large Language Model, the system effectively transforms lengthy video transcripts into structured, comprehensive, and easily digestible study notes. The architecture, built on React, Express.js, and the Google Gemini API, proves to be a robust and scalable solution for on-demand content processing. The primary contribution of this project lies in its NLP first approach to video content analysis. Unlike traditional summarizers that focus on visual keyframe extraction, S-Notevid prioritizes the semantic richness of the spoken transcript, making it an exceptionally effective tool for lectures, tutorials, and other knowledge-dense videos. This approach not only saves users significant time and effort but also provides a deeper, more organized understanding of the material.

The system's practical implementation addresses a genuine need in modern education and content consumption. With the exponential growth of video-based learning resources, tools that can efficiently extract and organize information become increasingly valuable. S-Notevid stands as a successful proof- of-concept for a new generation of AI-powered educational tools, effectively showcasing how modern API-driven AI can be applied to build practical, user-centric applications that meet the evolving needs of learners.

REFERENCES

[1] K. E. Nair, S. A. Johns, and A. John, "An overview of machine learning techniques applicable for summarisation of videos in education," in International Conference on Machine Learning and Cybernetics, Kollam, India: TKM College of Engineering, 2019.
[2] N. Anand, R. K. Koshariya, and V. Garg, "VidSum - Automated Video Summarization using Deep Learning," in International Conference on Computer Science and Engineering, Noida, India: Jaypee Institute of Information Technology, 2024.
[3] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," IEEE Transactions on Multimedia, vol. 21, no. 8, pp. 1949–1961, Aug. 2019.
[4] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive Sequence-Graph Network for Video Summarization," IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1234–1245, May 2021.
[5] S. Lal, S. Duggal, and I. Sreedevi, "Online Video Summarization: Predicting Future To Better Summarize Present," IEEE Transactions on Multimedia, vol. 23, pp. 567–576, April 2022.
[6] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural In- formation Processing Systems 30 (NeurIPS 2017), 2017, pp. 5998–6008.
[7] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 1877–1901.
[8] R. Nallapati, B. Zhou, C. Gulcehre, and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in Pro- ceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 280–290.
[9] M. Denil, D. Deniset, and N. Freitas, "Learning where to attend with deep architectures for image tracking," in International Conference on Machine Learning, 2014.
[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[11] A. Knobel, W. Mu¨ller, and O. Freimuth, "Secure OAuth 2.0 Implemen- tation Patterns," Journal of Web Security, vol. 15, no. 2, pp. 45–62, 2022.
[12] C. V. Ellis and S. M. Becker, "Efficient State Management in Single- Page Applications," in International Conference on Web Services, pp. 234–245, 2023.
[13] D. Krammer and B. Schneier, "Best Practices in API Security," ACM Transactions on Security and Privacy, vol. 18, no. 3, pp. 1–28, Sept. 2023.
[14] L. Richardson and S. Ruby, RESTful Web Services, O'Reilly Media, 2nd ed., 2017.
[15] A. Tanenbaum and M. van Steen, Distributed Systems: Principles and Paradigms, Prentice Hall, 3rd ed., 2017.
[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521,pp. 436–444, May 2015.
[17] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in International Conference on Learning Representations, 2021.
[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[20] C. Papineni, M. Roukos, W. Ward, and W. Zhu, "BLEU: Automatic Evaluation of Machine Translation," in Proceedings of the 40th Annual Meeting of ACL, 2002, pp. 311–318.
[21] X. Zhang and Y. LeCun, "Text Understanding from Scratch," in arXiv preprint arXiv:1502.01852, 2015.
[22] R. Socher, B. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in EMNLP, 2013.
[23] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks," in NeurIPS, 2015.
[24] C. Olah, "Understanding LSTM Networks," Blog post: colah.github.io, 2015.
[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in NIPS Deep Learning and Representation Learning Workshop, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY