EUROPEAN JOURNALS OF EMERGING COMPUTER VISION AND NATURAL LANGUAGE PROCESSING

Fusing Pixels and Prose: The Transformative Impact of Integrating Computer Vision and Natural Language Processing on Multimedia Robotics Applications

Dr. Nareen X. Veltrix
Department of Computer Science, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Dr. Tamos I. Drevik
School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA

V0LUME03 ISSUE01 (2026)

Published Date: 06 January 2026 // Page no.: - 01-08

ABSTRACT

Imagine a robot that doesn't just see the world, but can tell you about it. A machine that can follow your spoken instructions not just by rote, but with a genuine understanding of the objects and actions involved. This is the future being built at the intersection of two of artificial intelligence's most powerful fields: Computer Vision (CV) and Natural Language Processing (NLP). This article explores that exciting frontier. We'll journey through the core technologies that allow machines to see and to speak, from digital "eyes" that recognize objects to "minds" that can process language. Our main focus will be on the clever methods researchers use to weave these two abilities together, creating a shared understanding between pixels and prose. We'll look at the incredible results of this fusion: robots that can narrate a video, answer questions about a photograph, guide the visually impaired, or even learn a new task simply by watching and listening. Finally, we'll have an honest conversation about the tough puzzles that still need solving—like teaching machines true commonsense—and look ahead to a future where intelligent, collaborative robots become a beneficial part of our everyday lives.

Keywords: Computer Vision, Natural Language Processing (NLP), Multimedia Robotics, Human-Robot Interaction, Deep Learning, Multimodal Learning, Image Captioning, Video Description, Visual Question Answering (VQA), Semantic Grounding.

1. Introduction

For decades, robots were the stuff of assembly lines—powerful, precise, but ultimately blind and deaf to the nuanced world around them. Today, that is changing. We are in the midst of a quiet revolution, one that is giving robots the senses and the language to step out of the factory and into our lives. The catalysts for this transformation are two of the most exciting fields in artificial intelligence: Computer Vision (CV), the science of teaching computers to see, and Natural Language Processing (NLP), the art of teaching them to understand our language. While each field is a powerhouse on its own, their true magic is unleashed when they work together [3]. This partnership allows a machine to connect what it sees with what it can say, bridging the vast gap between raw pixels and real-world meaning.

1.1 The Pillars of Perception: Giving Robots Senses and Language

So, what does it mean for a computer to "see"? Computer Vision is about giving machines this very ability. Its goals can be thought of in terms of three big ideas: Recognition, Reconstruction, and Reorganization [6]. **Recognition** is about identifying the "what"—labeling the objects, people, and events in a scene. **Reconstruction** is about understanding the "where"—building a 3D map of the world from 2D images so a robot can navigate without bumping into things. And **Reorganization** is about finding the basic structure in an image—the edges and shapes that form the building blocks of vision [6]. We see these principles at work everywhere, from systems that spot unsafe behavior on construction sites [4] to the complex object detection that guides self-driving cars [11].

But seeing isn't enough. To truly collaborate with us, robots need to speak our language. That's where Natural Language Processing comes in. NLP is the key that unlocks communication between humans and machines [5]. It's a deeply complex challenge, involving everything from understanding the grammar of a sentence to grasping its underlying meaning and even the speaker's intent [9, 10]. Thanks to incredible advances in AI, we now have machines that can translate languages, understand our commands, and even detect emotion in our writing [1, 9]. For robotics, NLP means we can finally move beyond remote controls and complex programming to simply talking to our machines.

1.2 Bridging the Great Divide

The biggest challenge in bringing these two worlds together is what we call the **semantic gap**. Think about it: a robot's camera sees a grid of millions of colored pixels. A human, looking at the same scene, sees "a mother gently placing a blanket on her sleeping child." How do you get from one to the other? That chasm between pixels and meaning is the semantic gap [7]. Humans bridge it effortlessly, drawing on a lifetime of experience and conceptual understanding [7, 8]. For a robot to do the same, it must learn to **ground** the symbols of language—words like "blanket" and "sleeping"—in the visual patterns it perceives.

This means that integrating vision and language isn't just about running two programs at once. It's about creating a single, unified intelligence where seeing informs language, and language gives context to sight [13]. When you tell a robot, "Bring me the heavy blue mug," it needs to perform an incredible feat of multimodal reasoning. It must understand the command (NLP), then use its vision (CV) to find all the mugs, filter for the blue one, and maybe even guess which one looks "heavy" [14]. This ability is the foundation for a new generation of robotics, enabling everything from automatic video narration [15, 17, 18, 21, 22] to intuitive ways of teaching a robot a new skill [26].

As we imagine a future with robotic assistants in our homes, hospitals, and workplaces [2], their ability to see, understand, and communicate becomes essential. This article is a deep dive into this fascinating world. We'll explore the methods that make it all possible, celebrate the amazing applications already here, and honestly assess the challenges that still lie on the road ahead.

2. Methodologies for Vision-Language Integration

So, how do you actually teach a machine to connect words to images? It's one of the most fascinating puzzles in AI. The process isn't magic; it's a combination of clever architecture and vast amounts of data, relying on digital "brains" inspired by our own. Let's first look at the core tools for vision and language, and then explore the strategies for weaving them together.

2.1 Foundational Models and Architectures

2.1.1 Visual Feature Extraction: How a Machine Learns to See

The star player in modern computer vision is the **Convolutional Neural Network (CNN)**. Think about how you recognize a face. You don't process every pixel at once. Your brain first detects simple lines and curves, then combines them into features like eyes and a nose, and finally assembles those into a recognizable face. A CNN works in a remarkably similar way.

- It uses digital **filters** to scan an image, looking for basic patterns like edges and colors.
- As the image passes through layers of the network, these simple patterns are combined into more complex ones, like textures, shapes, and eventually, whole objects.
- Finally, the network takes this rich, high-level summary of the image and uses it to make a prediction, like "this is a cat."

In vision-language models, we often use a pre-trained CNN as our expert "eye." We show it an image and grab the final summary it produces—a compact numerical fingerprint of the image's content. This fingerprint, or "thought vector," is the visual information we'll pass on to the rest of the system [39].

2.1.2 Language Representation: Turning Words into Numbers

Machines don't understand words; they understand numbers. So, the first step in any NLP task is to translate language into a format a computer can work with.

- **Word Embeddings:** The real breakthrough here was the idea of **word embeddings** [40]. Imagine a giant map where every word has its own coordinates. On this map, words with similar meanings, like "happy" and "joyful," are placed close together. These embeddings are powerful because they capture the subtle relationships between words, allowing a machine to understand that "king" is to "queen" as "man" is to "woman" [25].
- **Sequence Modeling:** Of course, the meaning of a sentence depends on the order of its words. To capture this, we use models like **Recurrent Neural Networks (RNNs)**. An RNN reads a sentence one word at a time, keeping a running "memory" of what it has seen so far. More advanced versions, like LSTMs and GRUs [33], have more sophisticated memories, allowing them to understand the context of long, complex sentences and documents [34].

2.2 Core Integration Strategies

Once we have numerical representations of both what the robot sees and what it hears, the real work begins: connecting them.

2.2.1 Creating a Shared "Language"

The most direct approach is to create a shared multimodal "language" or embedding space. The goal is to train a system so that the numerical fingerprint of an image lands in the same spot on our map as the numerical fingerprint of its description. We do this by showing the model countless examples, rewarding it when it pulls correct image-sentence pairs together and penalizing it when it gets them wrong [24, 23]. More advanced techniques even learn to connect specific words or phrases, like "a brown dog," to the exact pixels in the image where that dog appears [28].

2.2.2 The Encoder-Decoder: A Translator for Vision

A very popular method, borrowed from machine translation, is the **encoder-decoder** architecture [17, 21].

- **The Encoder (The "Reader"):** A CNN acts as the encoder. It "reads" the image and condenses it down to its essential meaning, producing a single, rich context vector.
- **The Decoder (The "Writer"):** An RNN acts as the decoder. It takes this context vector and begins to "write" the description, generating the sentence one word at a time, using the memory of the words it's already written to inform its next choice.

2.2.3 Attention: Focusing on What Matters

A limitation of the basic encoder-decoder is that it's like trying to describe a complex painting after only a single glance. **Attention mechanisms** solve this. They allow the "writer" (the decoder) to look back at different parts of the image as it generates each word. When it's about to write the word "frisbee," the attention mechanism lets it focus intently on the part of the image containing the frisbee, resulting in far more accurate and detailed descriptions.

2.2.4 Multimodal Fusion: Thinking with Two Senses

For a task like Visual Question Answering (VQA), where the robot has to reason about an image to answer a question, it needs to truly fuse its two "senses." Simply sticking the visual and language data together isn't enough. Advanced techniques like **Multimodal Compact Bilinear Pooling (MCBP)** provide a way to intricately weave the two streams of information together, allowing the model to capture the complex relationships needed to answer the question correctly [38].

2.3 The Role of External Knowledge

Finally, researchers are realizing that to truly understand the world, robots need more than just the data they see and hear. They need a dose of commonsense.

- **Learning from Books:** By "reading" huge amounts of text, a model can learn general knowledge about the world—for instance, that "playing a guitar" often involves "strumming" and produces "music." This knowledge can then help it describe a video of someone playing guitar more richly [16, 17].
- **Using Knowledge Graphs:** A more structured approach is to give the robot access to a knowledge graph—a sort of digital encyclopedia of commonsense facts. The robot might see a <person, riding, bicycle> and use the knowledge graph to infer that a bicycle is a vehicle and riding is a form of transportation, allowing for much deeper reasoning [20].

These diverse strategies are the building blocks that allow us to create robots that can finally begin to bridge the gap between seeing and understanding.

3. Results and Applications

The fusion of computer vision and natural language processing is not just a theoretical exercise; it's producing tangible results and powering applications that were once the stuff of science fiction. By enabling robots to describe, interact with, and reason about their environment, this technology is moving from the lab into the real world.

3.1 Giving the World a Voice

One of the most stunning achievements of this field is giving machines the ability to describe what they see in their own words.

- **Image Captioning:** This is the foundational task: generating a sentence to describe a picture. We've moved far beyond clunky, template-based phrases. Modern systems, using the encoder-decoder and attention models we've discussed, can look at a photo and produce fluent, accurate descriptions like, "A young boy in a red jacket throws a frisbee in a park" [24, 28]. The quality is often so high that it's hard to distinguish from a caption written by a human [21].
- **Video Description:** Taking this a step further, video description requires understanding a scene as it unfolds over time. This is a much harder problem, but models can now generate concise summaries of video clips [17] or even weave together key moments into a short story [22]. The next great challenge is tackling long, unedited videos from the real world, with all their complexity and scene changes [18].
- **Dense Captioning and Scene Graphs:** For an even deeper understanding, dense captioning models can identify and describe many different things happening in a single image. A related and powerful idea is **scene graph generation**. Instead of a sentence, the model produces a structured map of the scene, like <man, wearing, hat> and <hat, on, head>. While less poetic, this format is perfect for a machine to reason with later

[20].

3.2 The Robot as a Collaborative Partner

Perhaps the most exciting application is the ability to interact with robots using natural language. This transforms them from tools we must program into partners we can instruct.

- **Instruction Following:** This is the holy grail of interactive robotics. Systems now exist that allow a robot to follow commands like, "Go down the hall and take the first door on your left," or "Pick up the red ball from the table" [14]. To do this, the robot must seamlessly blend its skills: it parses the command (NLP), visually identifies the objects and locations ("door," "table"), understands the spatial relationships ("on your left"), and then executes the correct action.

- **Learning by Watching and Listening:** This integration opens up incredibly intuitive ways to teach robots new things. Instead of writing code, you can simply show and tell. Imagine saying, "This is how you water the plant. You pick up the watering can like this, and pour gently at the base." The robot uses its vision to watch what you do and its language processing to understand your explanation, linking them together to learn the new skill in a rich, human-like way [26].

3.3 Answering Questions About the Visual World

Visual Question Answering (VQA) is a key benchmark for vision-language intelligence. You give the machine an image and ask a question, and it has to give you the right answer. This requires a much deeper level of reasoning than just describing a scene.

- **From Simple to Complex:** VQA questions can be as simple as "What color is the car?" or as complex as "Is the person in the photo likely to be happy?" Answering the latter requires a degree of commonsense reasoning that is still a major research challenge.

- **The Need for Fusion:** To succeed at VQA, a model must be an expert at fusing its senses. It has to understand the question, find the relevant parts of the image, and then intelligently combine that information to come up with an answer. This is where advanced techniques like Multimodal Compact Bilinear Pooling have been so important [38], and where large datasets for training have been indispensable [35, 36].

3.4 Specialized and Social Applications

This powerful combination of technologies is also finding its way into a growing number of specialized and socially beneficial roles.

- **Social Robotics:** For a robot to be an effective companion or assistant, it needs social skills. By combining what it sees (a person's facial expression) with what it hears (the tone of their voice), a social robot can better infer a person's emotional state and respond more empathetically, making for a much more natural and engaging interaction [1, 10].

- **Assistive Robotics for the Visually Impaired:** This is a truly life-changing application. Imagine a blind person wearing a pair of smart glasses. A camera captures their surroundings, and an AI model instantly identifies obstacles and objects of interest. The system then speaks to the user with simple, clear guidance: "Warning, curb ahead," or "Bicycle approaching on your left." This provides a layer of awareness and safety that goes far beyond what a traditional white cane can offer.

- **Smarter, Safer Workplaces:** In industrial settings, this technology can be a powerful safety tool. A vision system can monitor a construction site and spot unsafe behavior, like a worker not wearing a hard hat [4]. When it sees a problem, it can issue a clear, spoken warning. An inspection robot can survey a bridge, find a crack with its camera [2], and then automatically generate a detailed report, complete with a description, photo, and precise location of the damage.

These examples are just the beginning. They show that the fusion of vision and language is a deeply practical and transformative technology, paving the way for a future where robots are more intelligent, interactive, and helpful than ever before.

4. Discussion: Persistent Challenges and Future Directions

For all the incredible progress, the dream of a robot that can perceive and communicate with the fluidity of a human is still on the horizon. The path forward is filled with fascinating and formidable challenges. Having an honest conversation about these hurdles is the only way to navigate them and push the field toward its next breakthrough.

4.1 The Tough Problems We Still Need to Solve

4.1.1 The Grounding Problem: Do They Really Understand?

This is the big one. At its heart, the **grounding problem** asks whether our models truly *understand* the words they use, or if they are just incredibly sophisticated mimics [28]. A model can learn from millions of examples that the pixel patterns of a furry, four-legged animal are associated with the word "dog." But does it know what a dog *is*? Does it know they bark, chase balls, and are often called "man's best friend"? Mostly, no. This gap between statistical correlation and real-world understanding is why models sometimes make bizarre mistakes, like captioning a photo of a man on a skateboard as "A man is riding a horse on a city street." The shapes are vaguely similar, but the model lacks the basic commonsense to know this is impossible. True grounding is the next great leap [20, 28].

4.1.2 Bias, Fairness, and Unintended Consequences

AI models learn from the data we give them. When that data is scraped from the internet, it comes with all of our human biases baked in [13]. If a model sees that in photos, "cooking" is usually associated with women and "engineering" is usually associated with men, it will learn and reinforce that stereotype. A robot operating with these biases in the real world could be inequitable or even harmful. Tackling this is one of the most urgent tasks in AI, requiring not just better algorithms but a deep ethical commitment to building fair and responsible systems.

4.1.3 The Lego Problem: Compositional Generalization

Humans have an amazing ability to understand new things by combining concepts we already know. If you know what "purple" is and you know what a "giraffe" is, you can instantly imagine a "purple giraffe." This is called **compositional generalization**. Our AI models are surprisingly bad at this. They are great at recognizing things they've seen many times in their training data, but they often fail when presented with a novel combination. This makes them brittle and less adaptable to the ever-changing real world.

4.1.4 The Need for Speed: Real-Time Processing

The most powerful vision-language models are often computational behemoths, requiring massive server farms to run. This is a major problem for a robot that needs to make decisions in a split second. A self-driving car can't wait five seconds to identify a pedestrian. While clever solutions like edge computing can help [1], the race is on to create models that are both powerful and lightweight enough to run on the limited hardware inside a mobile robot.

4.1.5 The Ambiguity of It All

Both our language and our world are full of ambiguity. Does "bat" mean a piece of sports equipment or a flying mammal? Is that dark shape in the road a harmless shadow or a dangerous pothole? Humans use context to figure these things out almost instantly, but it's a huge challenge for machines [12]. A robot that can't handle ambiguity is a robot that can't be trusted in complex situations.

4.1.6 Are We Measuring the Right Thing?

How do we even know if a model is doing a good job? For image captioning, we often use metrics that compare a machine's caption to one written by a human. But what if the machine says, "A man is throwing a red disc," and the human wrote, "A boy is tossing a frisbee"? The machine's description is perfectly correct, but it would get a low score because the words don't match exactly. We need to develop smarter evaluation methods that can judge the *meaning* and *quality* of a response, not just the superficial words [21].

4.2 The Exciting Road Ahead

These challenges aren't dead ends; they are signposts pointing toward the future of research.

- **From Correlation to Causation:** The next frontier is to build models that understand cause and effect. A robot shouldn't just learn that wet streets are correlated with clouds; it should understand that clouds *cause* rain, which *causes* the streets to get wet. This leap from correlation to causal reasoning is a key step toward true intelligence.

- **The Curious Robot: Interactive and Lifelong Learning:** Instead of being trained just once, future robots should be like curious toddlers, constantly learning from their interactions with the world. They should be able to ask questions when they're confused ("Did you mean this cup or that one?") and update their knowledge without forgetting everything they've learned before [26].

- **The Best of Both Worlds: Neuro-Symbolic Models:** A very promising path is to build hybrid models that combine the pattern-recognition power of neural networks with the logical reasoning of classical AI. The idea is to have the "neuro" part handle perception (seeing the scene) and the "symbolic" part handle reasoning (thinking logically about what it sees) [20].

- **Learning by Doing: Embodiment:** Finally, there's a growing understanding that true intelligence is embodied. A robot will understand the word "push" far more deeply if it can physically push objects and see what happens, rather than just reading the word in a book. The future of robotics is one where learning is an active, physical process.

The journey to fuse pixels and prose is just beginning. The challenges are immense, but the potential rewards—a future of more intelligent, helpful, and collaborative robots—are more than worth the effort.

5. Conclusion

The marriage of computer vision and natural language processing represents a defining chapter in the story of artificial intelligence. This powerful partnership is fundamentally reshaping what is possible, transforming machines from simple tools into active participants in our world—agents that can see, act, and, crucially, communicate. We have journeyed through the core concepts, from the neural networks that act as digital eyes to the complex architectures that weave sight and sound into a single, coherent understanding.

We've seen the incredible applications this synergy makes possible: robots that can narrate a scene, follow our spoken commands, and assist us in our daily lives. The potential to positively impact humanity is undeniable, with real-world applications emerging in assistive technology, industrial safety, and social robotics that promise to make our world safer, more accessible, and more efficient.

Yet, we have also looked clear-eyed at the mountain we still have to climb. The path to a machine with true, human-like understanding is steep, marked by profound challenges like the semantic grounding problem, the danger of algorithmic bias, and the need for genuine commonsense reasoning. These are not small hurdles to be easily overcome; they are fundamental questions that force us to think deeply about the nature of intelligence itself.

The future of this field is bright and points toward systems that are more interactive, more curious, and more deeply connected to the physical world. The pursuit of robots that can learn continuously through conversation and action, and the development of hybrid models that combine the best of different AI philosophies, will continue to push the boundaries of innovation. Ultimately, fusing pixels and prose is more than an engineering challenge. It is a critical and inspiring step toward the long-held dream of creating truly intelligent machines—not as replacements for humanity, but as capable and collaborative partners, ready to help us build a better future.

References

[1] G. Yin, Intelligent framework for social robots based on artificial intelligence-driven mobile edge computing, Computers & Electrical Engineering, 96, Part B, (2021).

[2] Fisher, M., Cardoso, R. C., Collins, E. C., Dadswell, C., Dennis, L. A., Dixon, C., ... & Webster, M., An overview of verification and validation challenges for inspection robots, Robotics, 10, 67 (2021).

[3] A. Jamshed and M. M. Fraz, NLP Meets Vision for Visual Interpretation - A Retrospective Insight and Future directions, 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), 1-8 (2021).

[4] W. Fang, P. E.D. Love, H. Luo, L. Ding, Computer vision for behaviour-based safety in construction: A review and future directions, Advanced Engineering Informatics, 43, (2020).

[5] H. Sharma, Improving Natural Language Processing tasks by Using Machine Learning Techniques, 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 1-5 (2021).

[6] M. Jitendra, P. Arbeláez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani, The three R's of computer vision: Recognition, reconstruction and reorganization, Pattern Recognition Letters, 72, 4-14 (2016).

[7] P. Gärdenfors, The Geometry of Meaning: Semantics Based on Conceptual Spaces, MIT Press, (2014).

[8] E. Dockrell, D. Messer, R. George, and A. Ralli, Beyond naming patterns in children with WFDs—Definitions for nouns and verbs, Journal of Neurolinguistics, 16, 191-211 (2003).

[9] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, Natural language processing advancements by deep learning: A survey, arXiv preprint arXiv:2003.01200 (2020).

[10] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, and C. Santos-Libarino, Emotion detection for social robots based on nlp transformers and an emotion ontology, Sensors, 21, 1322 (2021).

[11] S., Zhenfeng, W. Wu, Z. Wang, W. Du, and C. Li, Seaships: A large-scale precisely annotated dataset for ship detection, IEEE transactions on multimedia, 20, 2593-2604 (2018).

[12] https://monkeylearn.com/blog/natural-language-processing-challenges/ , last vist 1/2/2022.

[13] C. Zhang, Z. Yang, X. He and L. Deng, Multimodal Intelligence: Representation Learning, Information Fusion, and Applications, in IEEE Journal of Selected Topics in Signal Processing, 14, 478-493 (2020).

[14] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy, Understanding natural language commands for robotic navigation and mobile manipulation. In Proceedings of the AAAI Conference on Artificial Intelligence, 25, 1507-1514 (2011).

[15] Y. Yezhou, C. Teo, H. Daumé III, and Y. Aloimonos, Corpus-guided sentence generation of natural images, In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 444-454 (2011).

[16] T. S. Motwani, R. J. Mooney, Improving Video Activity Recognition using Object Recognition and Text Mining, In

Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012), 600-605 (2012).

[17] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko and S. Guadarrama, Generating Natural-Language Video Descriptions Using Text-Mined Knowledge, In Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI-2013), 541-547 (2013).

[18] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R Mooney, Integrating language and vision to generate natural language descriptions of videos in the wild, Proceedings of the 25th International Conference on Computational Linguistics (COLING), (2014).

[19] Y. Yezhou, C. L. Teo, C. Fermüller, and Y. Aloimonos, Robots with language: Multi-label visual recognition using NLP, In IEEE International Conference on Robotics and Automation, 4256-4262 (2013).

[20] S. Aditya, Y. Yang, C. Baral, C. Fermuller, and Y. Aloimonos, From images to sentences through scene description graphs using commonsense reasoning and knowledge, arXiv preprint arXiv, (2015).

[21] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. I. Cinbis, F. Keller, A. Muscat, and B. Plank, Automatic description generation from images: A survey of models, datasets, and evaluation measures, Journal of Artificial Intelligence Research, 55, 409-442 (2016).

[22] P. Das, C. Xu, R. Doell, and J. Corso, A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2634-264 (2013).

[23] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, International journal of computer vision, 106, 210-233 (2014).

[24] A. Karpathy and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3128-3137 (2015).

[25] R. Schwartz, R. Reichart and A. Rappoport, Symmetric pattern based word embeddings for improved word similarity prediction, In CoNLL, 2015, 258-267 (2015).

[26] N. Shukla, C. Xiong, and S. C. Zhu, A unified framework for human-robot knowledge transfer, In Proceedings of the 2015 AAAI Fall Symposium Series, (2015).

[27] Carina Silberer, Vittorio Ferrari, and Mirella Lapat, Models of semantic representation with visual attributes, In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 572-582 (2013).

[28] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2, 207-218 (2014).

[29] M. Tapaswi, M. B¨auml, and R. Stiefelhagen, Book2movie: Aligning video scenes with book chapters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1827–1835 (2015).

[30] I. Abdalla Mohamed, A. Ben Aissa, L. F. Hussein, Ahmed I. Taloba, and T. kallel, A new model for epidemic prediction: COVID-19 in kingdom saudi arabia case study", Materials Today: Proceedings, (2021).

[31] Ahmed. I. Taloba and S. S. I. Ismail, An Intelligent Hybrid Technique of Decision Tree and Genetic Algorithm for E-Mail Spam Detection, Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 99-104 (2019).

[32] Ahmed I. Taloba, M. R. Riad and T. H. A. Soliman, Developing an efficient spectral clustering algorithm on large scale graphs in spark, Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), 292-298 (2017).

[33] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1422–1432.

[34] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in International conference on machine learning, 2014, pp. 1188–1196.

[35] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," ArXiv Prepr. ArXiv150805326, 2015.

[36] Y. Yang, W. Yih, and C. Meek, "Wikiqa: A challenge dataset for open-domain question answering," in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2013–2018.

[37] W. Yih, M. Richardson, C. Meek, M.-W. Chang, and J. Suh, "The value of semantic parse labeling for knowledge base question answering," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 201–206.

[38] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and

M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” ArXiv Prepr. ArXiv160601847, 2016.

[39] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” ArXiv Prepr. ArXiv14042188, 2014.

[40] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Adv. Neural Inf. Process. Syst., vol. 26, 2013.