SYNERGIES OF SIGHT AND LANGUAGE: A JOURNEY INTO HOW MACHINES LEARN TO SEE AND SPEAK

Dr. Leona V. Merake
Department of Cognitive Robotics, Delft University of Technology, Delft, Netherlands

Dr. Arvind S. Tomura
Artificial Intelligence and Vision Lab, University of Tokyo, Tokyo, Japan

V0LUME01 ISSUE01 (2024)

Published Date: 27 December 2024 // Page no.: - 93-100

ABSTRACT

For decades, we've dreamed of creating machines that can see the world as we do and talk to us about it. This once-distant dream is now a vibrant reality at the intersection of two of artificial intelligence's most ambitious fields: Computer Vision (CV) and Natural Language Processing (NLP). This article takes you on a comprehensive journey through this exciting landscape, exploring how we're teaching computers to connect pixels to prose. We'll start by exploring the fundamental questions that drive this field, like the "symbol grounding problem"—the puzzle of how words get their meaning—and use frameworks like Bloom's Taxonomy to map out what it truly means for a machine to "understand." We'll break down the core tasks of vision into the "3Rs" (Recognition, Reconstruction, Reorganization) and language into its essential layers (Syntax, Semantics, Pragmatics) to see exactly where these two worlds meet.

The heart of this survey is a deep dive into the toolbox of modern AI. We'll explore how machines learn to represent the world, from the early days of handcrafted visual features to the powerful deep learning models that create today's "image embeddings" and "word embeddings." From there, we'll investigate the ingenious architectures that fuse these senses together, including shared embedding spaces, the elegant encoder-decoder models that power image captioning, the clever attention mechanisms that let models "focus" on what's important, and the modular networks that allow for compositional reasoning.

We'll see these methods in action as we examine the key battlegrounds where progress is measured: tasks like generating captions for images and videos, answering complex questions about a scene (VQA), and retrieving images with a simple text query. We'll then venture into the world of robotics, where this technology is giving machines the ability to follow our commands, learn by watching us, and engage in grounded, meaningful dialogue. By weaving together insights from landmark studies, we'll paint a picture of the field's incredible achievements. Finally, we'll have a candid discussion about the tough challenges that lie ahead—the need for genuine commonsense, the subtle biases that creep into our data, and the quest for true understanding—as we look toward a future of embodied, communicative AI that promises to change our relationship with technology forever.

Keywords: Computer Vision, Natural Language Processing, Vision and Language, Multimodal Learning, Deep Learning, Image Captioning, Visual Question Answering (VQA), Robotics, Symbol Grounding, Human-Robot Interaction, Representation Learning, Distributional Semantics, Embodied AI.

INTRODUCTION

The Quest for a Seeing, Talking Machine

The ability to glance at a scene—a bustling market, a quiet park, a child's birthday party—and effortlessly describe it in words is a profoundly human skill. It's a seamless dance between our eyes and our minds, a cornerstone of how we share experiences and build understanding [146, 178]. For decades, researchers in Artificial Intelligence have been on a quest to bestow this same gift upon machines. This survey charts the course of that adventure, tracing the convergence of two once-separate disciplines—Computer Vision and Natural Language Processing—into a unified field that is pushing the very boundaries of what machines can do.

1.1. Two Worlds, One Goal: The Foundations of Vision and Language

To appreciate how we're bridging the gap between sight and speech, we first have to understand the worlds they come from.

1.1.1. What Does it Mean to "See"? The Three "Rs" of Vision

Computer vision, at its core, is about turning light into understanding. This complex process can be thought of in terms of three fundamental goals, the "3Rs" [185]:

● Reconstruction: This is the magician's trick of conjuring a 3D world from a flat, 2D image. It involves figuring out the shape, layout, and geometry of a scene from clues like shading, perspective, and motion [25]. Language helps here by giving us hints; knowing you're looking at a "car" gives the system a powerful clue about the 3D shape it should expect to find.

- Recognition: This is the task we're most familiar with—putting names to faces, objects, and places. It's about assigning meaningful labels to the world, from identifying a "dog" in a photo to classifying an entire scene as a "beach" [97, 217, 261]. Recognition is the most natural handshake between vision and language, as its output is a stream of words.

- Reorganization: Before we can recognize anything, we need to make sense of the chaotic mess of pixels. Reorganization is the process of grouping those pixels into meaningful wholes—finding the edges, separating objects from their background, and clustering textures [190, 274]. It's the visual brain's way of tidying up the world so that recognition can happen.

These three tasks don't work in isolation; they're in a constant, collaborative dance. The neat outlines from reorganization help with recognition, and the labels from recognition help guide the 3D reconstruction.

1.1.2. What Does it Mean to "Speak"? The Layers of Language

In the world of language, the journey from thought to utterance can be mapped using the Vauquois triangle, a model from machine translation that shows the different layers of analysis needed to understand and generate language [286]. To "translate" from a perception to a description involves navigating these layers of meaning:

- Syntax: These are the rules of the road for language—its grammar. Syntax ensures that sentences are put together in a way that makes sense. A vision-language model must master syntax to avoid producing gibberish.

- Semantics: This is the heart of language—its meaning. It's about knowing that "dog bites man" means something terrifyingly different from "man bites dog," even though they use the same words. It's understanding that a "red ball" is an object that has both the property of being "red" and being a "ball."

- Pragmatics: This is the subtle, social layer of language. It's the ability to read between the lines and understand meaning in context. If a robot drops a tray of glasses and you say, "Great job," pragmatics is what allows it to understand your sarcasm and not take it as a compliment.

1.2. The Elephant in the Room: The Symbol Grounding Problem

The biggest philosophical and technical hurdle in uniting these two worlds is the symbol grounding problem [138]. It's a simple but profound question: how do words get their meaning? A computer that has only ever read text can learn that the word "cat" often appears near the word "purr," but it has no idea what a cat is. Its knowledge is a closed loop of symbols, unattached to reality.

To truly understand, the symbols must be grounded in the real world—in perceptual experience. This means:

- Connecting the word "cat" to the sight, sound, and feel of an actual cat (Physical Grounding).

- Tracking that specific cat as it moves around a room (Perceptual Anchoring) [73].

- Knowing that the meaning of "chair" isn't just its visual shape, but the fact that you can sit on it (Grounding in Action) [65, 239].

Without this grounding, a machine is just a clever parrot, repeating patterns without comprehension.

1.3. How Do We Know if It's Working? A Roadmap for Understanding

So how do we measure progress on this grand quest? We can borrow a framework from the world of education: Bloom's Taxonomy [43, 257]. It provides a ladder of cognitive skills, from simple to complex, that gives us a roadmap for building and testing smarter machines.

1. Knowledge: Can the machine recall basic facts? (What is in the image?)

2. Comprehension: Can it explain or summarize? (Why is that person holding an umbrella?)

3. Application: Can it use what it knows in a new situation? (How do I pick up that cup?)

4. Analysis: Can it break down information and see relationships? (How is this kitchen different from the last one?)

5. Synthesis: Can it create something new? (Plan a path to get the apple from the counter.)

6. Evaluation: Can it make judgments? (Which of these two routes is safer?)

Today's AI is getting remarkably good at the first few rungs of this ladder. The highest rungs—true analysis, synthesis, and evaluation—remain the formidable, exciting challenges for the future. This survey will explore the methods that have gotten us this far and light the path toward what's next.

2. The AI Toolbox: How Machines Learn to See and Speak

At the heart of the vision-language revolution is a set of powerful tools and techniques for turning raw data into meaningful understanding. This process starts with learning good representations for both images and text, and then building clever architectures to fuse them together.

2.1. Learning the Language of Vision and Text

Before a machine can reason about the world, it needs to convert the messy, unstructured data of pixels and characters into a clean, mathematical format it can work with: vectors, or what we call embeddings.

2.1.1. Visual Representations: From Pixels to Perception

The way we represent images has evolved dramatically over the years.

- The Old School: Handcrafted Features: In the early days, computer vision experts would act like digital artisans, carefully designing algorithms to find interesting patterns in images. Tools like SIFT [183] and SURF [26, 27] were brilliant at finding stable "keypoints" in an image—like the corner of an eye or the petal of a flower—that could be recognized even if the image was rotated or resized.

- The "Bag-of-Words" Idea: To describe a whole image, researchers borrowed an idea from linguistics: the Bag-of-Visual-Words (BoVW) model [85]. Imagine taking all the keypoints from thousands of images and clustering them into a "visual vocabulary." Any new image could then be described by counting how many times each "visual word" appeared. It was a powerful technique, but it threw away all spatial information—it knew an image contained a nose, an eye, and a mouth, but not that they were arranged into a face.

- The Deep Learning Revolution: Image Embeddings: Everything changed with the arrival of Deep Convolutional Neural Networks (CNNs) [172]. Inspired by the human visual cortex, these networks, with names like AlexNet [166], VGGNet [252], and ResNet [141], learn to recognize features automatically by being trained on millions of images from datasets like ImageNet [35]. The final layer of a trained CNN produces a compact vector—an image embedding—that acts as a rich, semantic fingerprint of the image. These embeddings have become the universal language for describing images in modern AI.

- Beyond 2D: Seeing in 3D: For robots that need to navigate the real world, a flat image isn't enough. They need to see in 3D. Using data from sensors like LiDAR or depth cameras, we can represent the world as a point cloud or a grid of voxels. New kinds of deep learning models, like 3D ShapeNets [300], are now learning to interpret this 3D data directly, giving robots a much richer understanding of the space around them [113].

2.1.2. Language Representations: The Magic of Word Embeddings

The world of text processing has had its own revolution.

- From Counts to Context: Distributional Semantics: The big idea here is that you can understand a word by the company it keeps [139]. Early models like LSA [83] and LDA [42] used statistical methods to find these relationships. But the real breakthrough came with models like Word2Vec [199] and GloVe [222]. These algorithms learn a dense vector—a word embedding—for every word. The magic is that the geometry of this new vector space captures meaning. The vector for "king" minus "man" plus "woman" results in a vector incredibly close to "queen." It's like teaching a computer to understand analogies.

- Understanding Sentences with RNNs: To understand a whole sentence, we need to process words in order. Recurrent Neural Networks (RNNs) do just this, reading a sentence one word at a time while maintaining a "memory" in their hidden state. However, standard RNNs have a short memory. That's why we now use more sophisticated versions like Long Short-Term Memory (LSTM) networks [142] and Gated Recurrent Units (GRUs) [69]. Their special "gate" mechanisms allow them to remember important information over much longer sequences, making them perfect for understanding paragraphs or even entire documents.

- Adding Structure with Semantic Parsing: Sometimes, an embedding isn't enough. For a robot to follow a command, it needs a precise, unambiguous instruction. Semantic Parsing is the process of translating a natural language sentence into a formal, logical representation that a computer can execute [323, 324]. It's the AI equivalent of a lawyer drafting a contract, leaving no room for misinterpretation.

2.2. Architectures of Integration: Fusing the Senses

Once we have these powerful vector representations, the next challenge is to build architectures that can intelligently combine them.

2.2.1. Finding Common Ground: Joint Embedding Models

The most straightforward approach is to create a "multimodal" space where vision and language can live together. The goal is to learn a mapping so that the vector for an image of a sunset is very close to the vector for the sentence "a beautiful sunset over the ocean" [158, 256]. Once this shared space is learned, you can perform powerful cross-modal retrieval—finding images with text queries and vice versa.

2.2.2. The Storyteller: Encoder-Decoder Models

For tasks like image captioning, where the goal is to generate text, the encoder-decoder framework is king [298].

- The Encoder (a CNN) "looks" at the image and summarizes it into a single thought vector.

- The Decoder (an LSTM) takes this thought vector and begins to "speak," generating the caption one word at a time, just like a human telling a story.

2.2.3. Paying Attention: The Focus Mechanism

A limitation of the basic encoder-decoder model is that it forces the entire, rich image into one tiny thought vector—a huge information bottleneck. The attention mechanism was a brilliant solution to this problem [22, 317]. Instead of looking at the whole image at once, it allows the decoder to "focus" on different parts of the image as it generates each word. When it's about to say "dog," it pays attention to the dog. When it's about to say "ball," it shifts its focus to the ball. This simple idea dramatically improved the quality and detail of generated captions.

2.2.4. Answering Questions: Advanced Fusion

Answering a question about an image requires a more sophisticated fusion of the two modalities.

- Co-Attention: These models build on the idea of attention by allowing the image and question to attend to each other. The model uses the question to find relevant parts of the image, and uses the image to find the most important words in the question, creating a virtuous cycle of reasoning.
- Modular Networks: Recognizing that questions are often compositional, Neural Module Networks [8, 9] are like building with LEGOs. The system learns a set of basic modules (like find, count, query_color) and learns how to snap them together in the right order based on the question's structure.
- Memory Networks: For really complex reasoning that might take multiple steps, Memory Networks provide the AI with a kind of scratchpad [255, 315]. They can store the intermediate results of their thought process, allowing them to tackle much harder problems that require chaining facts together.

3. The Proving Grounds: Core Tasks and Applications

The methods we've discussed aren't just theoretical; they are being put to the test in a range of challenging tasks that serve as the benchmarks for the entire field. The creation of large, public datasets has been the fuel for this competitive and collaborative progress.

3.1. Describing the World in Detail

3.1.1. Beyond Labels: Attribute-Based Vision

Instead of just slapping a single label like "bird" on an object, attribute-based vision seeks to describe it with a rich vocabulary of properties: has_wings, is_small, has_a_yellow_beak [92, 171]. This fine-grained approach is incredibly powerful. It allows for Zero-Shot Learning, where a machine can learn to recognize an "oriole" simply by being told its attributes, even if it has never seen one before [171]. It also enables Relative Attributes, allowing the machine to make nuanced comparisons, like judging that one dog is "fluffier" than another [221].

3.1.2. Understanding Actions: Human-Object Interaction (HOI)

This area moves from recognizing what things are to understanding what things are doing. The goal is to identify <human, verb, object> triplets, like <person, riding, bicycle> [53]. The next level of this is Visual Semantic Role Labeling, where the system parses the entire visual event, identifying not just the action ("cutting"), but also the agent (the person), the patient (the thing being cut), and the instrument (the knife) [135, 321]. It's about understanding the story of an action.

3.2. Generating a Visual Narrative

3.2.1. Image Captioning: A Picture is Worth a Thousand Words

This is the quintessential vision-language task: teaching a machine to write a sentence describing an image. The field has moved at lightning speed from clunky, template-based systems [161] to the fluid, often poetic, sentences generated by today's attention-based models [298, 317]. This progress has been benchmarked on datasets like Flickr8k, Flickr30k, and the massive Microsoft COCO dataset [63], which provides a gold standard of multiple human captions for every image.

3.2.2. Video Captioning: Describing the Moving World

Video adds the complexity of time. A video captioning system must not only recognize objects and actions but also weave them into a coherent narrative. Models now use sequences of image features fed into LSTMs [296, 297], and even more advanced techniques like 3D CNNs to understand motion and hierarchical RNNs to generate entire paragraphs that tell the story of a video clip [320].

3.3. The Ultimate Test: Visual Question Answering (VQA)

VQA is perhaps the most comprehensive test of a system's understanding. It requires the machine to answer a specific, free-form question about an image [11]. This is far harder than captioning because it demands targeted reasoning. Questions can range from simple ("What color is the bus?") to complex ("How many people are wearing glasses?"). The VQA dataset [11] and the incredibly rich Visual Genome dataset [165] are the primary arenas for this task. A major hurdle has been overcoming dataset biases, where models learn to "cheat" by memorizing common answers rather than truly reasoning about the image.

3.4. Bringing it to Life: Applications in Robotics

This is where the virtual meets the physical. The fusion of vision and language is what will allow robots to leave the factory floor and enter our homes and lives.

- Instruction Following: The goal is to build a robot that can understand our commands, like "Please bring me the blue mug from the kitchen counter" [279]. This requires sophisticated semantic parsing to break the command down into a series of find, go, and grasp actions.
- Learning from Demonstration (LfD): Robots can learn new skills simply by watching us [12]. Language provides a powerful shortcut. If a person says, "Now I'm going to carefully wipe the spill," it gives the robot crucial information about the goal and constraints of the action it's observing.
- Situated Human-Robot Dialogue: For a robot to be a true collaborator, it needs to be able to talk with us about

the world we share [192]. This means being able to understand an ambiguous command like "hand me that one" by looking at where we're pointing, and being smart enough to ask, "Do you mean the big one or the small one?" when it's confused [278]. This is the frontier of embodied, interactive AI.

4. The Road Ahead: Grand Challenges and Future Dreams

For all the breathtaking progress, our journey toward a truly seeing, talking machine has only just begun. Today's models are brilliant in many ways, but their understanding is often shallow and brittle. The road ahead is filled with fascinating challenges that will require new ideas and a deeper connection to the principles of human cognition.

4.1. The Hurdles We Still Face

- The Commonsense Gap: Our models can describe what is in a picture with startling accuracy, but they often have no idea why. They can see a person putting a turkey in an oven, but they can't infer that it's Thanksgiving, that the person intends to cook a meal, or that the oven will soon be hot. This deep, commonsense reasoning is the missing ingredient in today's AI [1, 3].

- The Bias Trap: AI models are masters of finding patterns, but this is a double-edged sword. They are notoriously good at picking up on and exploiting subtle biases in the data they're trained on [283]. A VQA model might learn that questions about tennis are often answered with "tennis racket" without ever really understanding the sport. This makes them unreliable and prone to embarrassing mistakes when faced with the real world.

- The Compositionality Problem: Humans can effortlessly understand new combinations of things they already know—a "purple banana" is weird, but we know exactly what it means. AI struggles with this. It's great at recognizing things it has seen before, but it often fails when asked to generalize to novel compositions.

- The Measurement Problem: How do we even measure success? Our automated metrics for tasks like image captioning are crude tools. They can tell us if the words are similar to a human's, but they can't tell us if the caption is factually correct, witty, or just plain nonsensical [94].

- The Grounding Problem Revisited: We've made progress in linking words to pixels, but have we truly solved the symbol grounding problem [138]? Many would argue that real understanding can't be learned from a static dataset of images, no matter how large. True grounding may require embodiment—the ability to move through the world, interact with it, and learn from the consequences of one's own actions [214].

4.2. Charting the Future: Where Do We Go From Here?

These challenges aren't dead ends; they are signposts pointing toward the most exciting frontiers of AI research.

1. From Correlation to Causation: The next great leap will be to move beyond recognizing patterns to understanding cause and effect. We need to build models that have an intuitive grasp of physics and causality, that can reason about why a glass fell and predict that it will shatter.

2. Understanding Stories, Not Just Snapshots: The focus is shifting from single images to understanding entire events and narratives [236, 263]. This means building models that can track the goals, plans, and relationships of characters over time, weaving individual moments into a coherent story.

3. Learning Through Interaction: The future of AI is not in passively training on static datasets. It's in active, lifelong learning. The most powerful systems will be embodied agents—robots—that can explore the world, ask questions, run experiments, and learn from a continuous stream of interactive experience [280].

4. The Best of Both Worlds: Neuro-Symbolic AI: To get the best of both deep learning's perceptual power and classical AI's logical reasoning, researchers are building hybrid neuro-symbolic models. The idea is to create systems that can both see the world and reason about it in a structured, transparent way.

5. Opening the Black Box: Explainable AI (XAI): If we're going to trust AI in our cars and our hospitals, it's not enough for it to be right; we need to know why it's right. Building models that can explain their decisions in natural language ("I'm turning left because your instruction was to go to the pharmacy, which I've identified on this corner") is no longer a luxury—it's a necessity.

CONCLUSION

The convergence of vision and language is more than just a fascinating technical problem; it's a quest to replicate one of the most essential aspects of human intelligence. Powered by deep learning and vast datasets, we've built systems that can describe the world in sentences, answer our questions, and take the first tentative steps toward interacting with us in our own language. The road ahead is long, and the deepest challenges of reasoning, generalization, and true, grounded understanding still await us. But as we continue to explore this synergy, guided by the twin lights of cognitive science and embodied robotics, we are not just building smarter machines. We are gaining a new, profound appreciation for the incredible computational marvel that is the human mind.

REFERENCES

1. Aditya, S., Yang, Y., Baral, C., Fermuller, C., & Aloimonos, Y. (2015). From images to sentences through scene description graphs using commonsense reasoning and knowledge. arXiv preprint arXiv:1511.03292.

2. Aksoy, E. E., Abramov, A., Dorr, J., Ning, K., Dellen, B., & Worgotter, F. (2011). Learning the semantics of object–action relations by observation. The International Journal of Robotics Research, 30(10), 1229-1249.

3. Aloimonos, Y., & Fermuller, C. (2015). The cognitive dialogue: A new model for vision implementing common sense reasoning. Image and Vision Computing, 34, 42-44.

4. Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., & Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. Journal of Machine Learning Research, 15(1), 2773-2832.

5. Anandkumar, A., Hsu, D., & Kakade, S. M. (2012a). A method of moments for mixture models and hidden Markov models. In Conference on Learning Theory.

6. Anandkumar, A., Liu, Y. K., Hsu, D. J., Foster, D. P., & Kakade, S. M. (2012b). A spectral algorithm for latent Dirichlet allocation. In Advances in Neural Information Processing Systems (pp. 917-925).

7. Anderson, A. J., Bruni, E., Bordignon, U., Poesio, M., & Baroni, M. (2013). Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1960-1970).

8. Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016a). Learning to compose neural networks for question answering. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

9. Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016b). Neural module networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 39-48).

10. Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. Psychological Review, 116(3), 463.

11. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425-2433).

12. Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. Robotics and Autonomous Systems, 57(5), 469-483.

13. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

14. Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley framenet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1 (pp. 86-90).

15. Bakir, G. H. (2007). Predicting structured data. MIT press.

16. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... & Schneider, N. (2012). Abstract meaning representation (AMR) 1.0 specification. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

17. Bandura, A. (1974). Psychological Modeling: Conflicting Theories. Transaction Publishers.

18. Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S., Fidler, S., ... & Salvi, D. (2012a). Video in sentences out. In UAI 2012.

19. Barbu, A., Michaux, A., Narayanaswamy, S., & Siskind, J. M. (2012b). Simultaneous object detection, tracking, and event recognition. In ACS 2012.

20. Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. Journal of Machine Learning Research, 3, 1107–1135.

21. Barnard, K., & Forsyth, D. (2001). Learning the semantics of words and pictures. In Proceedings of the 8th IEEE International Conference on Computer Vision, 2001 (ICCV 2001), Vol. 2. IEEE, 408–415.

22. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

23. Baroni, M. (2016). Grounding distributional semantics in the visual world. Language and Linguistics Compass, 10(1), 3–13.

24. Barranco, F., Fermuller, C., & Aloimonos, Y. (2014). Contour motion estimation for asynchronous event-driven cameras. Proc. IEEE, 102(10), 1537–1556.

25. Barron, J. T., & Malik, J. (2015). Shape, illumination, and reflectance from shading. IEEE Trans. Pattern Anal. Mach. Intell., 37(8), 1670–1687.

26. Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). Comput. Vis. Image Understand., 110(3), 346–359.

27. Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In Computer Vision–ECCV 2006. Springer, 404–417.

28. Banarescu, L., et al. (2012). Abstract meaning representation (AMR) 1.0 specification. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

29. Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell., 19(7), 711–720.

30. Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neur. Comput., 15(6), 1373–1396.

31. Beltagy, I., Roller, S., Cheng, P., Erk, K., & Mooney, R. J. (2015). Representing meaning with a combination of logical form and vectors. arXiv preprint arXiv:1505.06816.

32. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell., 35(8), 1798–1828.

33. Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. J. Mach. Learn. Res., 3, 1137–1155.

34. Bengio, Y., Larochelle, H., Lamblin, P., Popovici, D., Courville, A., Simard, C., ... & Erhan, D. (2007). Deep architectures for baby AI.

35. Berg, A., Deng, J., & Fei-Fei, L. (2010). Large scale visual recognition challenge (ILSVRC), 2010.

36. Berg, T. L., & Berg, A. C. (2009). Finding iconic images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009 (CVPR Workshops 2009). IEEE, 1–8.

37. Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y. W., ... & Forsyth, D. A. (2004). Names and faces in the news. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’04), Vol. 2. IEEE, II–848.

38. Berg, T. L., Berg, A. C., & Shih, J. (2010). Automatic attribute discovery and characterization from noisy web data. In Computer Vision–ECCV 2010. Springer, 663–676.

39. Berg, T. L., Forsyth, D., & others. (2006). Animals on the web. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. IEEE, 1463–1470.

40. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., ... & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. J. Artif. Intell. Res., 55, 409–442.

41. Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. ACM, 127–134.

42. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. J. Mach. Learn. Res., 3, 993–1022.

43. Bloom, B. S., & others. (1956). Taxonomy of educational objectives. Vol. 1: Cognitive domain. McKay, New York, NY, 20–24.

44. Bronstein, A. M., Bronstein, M. M., & Kimmel, R. (2005). Three-dimensional face recognition. Int. J. Comput. Vis., 64(1), 5–30.

45. Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in technicolor. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 136–145.

46. Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. J. Artif. Intell. Res., 49, 1–47.

47. Byron, D., Koller, A., Oberlander, J., Stoia, L., & Striegnitz, K. (2007). Generating instructions in virtual environments (GIVE): A challenge and an evaluation testbed for NLG.

48. Cangelosi, A. (2006). The grounding and sharing of symbols. Pragm. Cogn., 14(2), 275–285.

49. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In AAAI, Vol. 5. 3.

50. Carrasco, M. (2011). Visual attention: The past 25 years. Vis. Res., 51(13), 1484–1525.

51. Carreira, J., & Sminchisescu, C. (2010). Constrained parametric min-cuts for automatic object segmentation. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 3241–3248.

52. Chang, A. X., Savva, M., & Manning, C. D. (2014). Semantic parsing for text to 3d scene generation. ACL 2014, 17.

53. Chao, Y. W., Wang, Z., He, Y., Wang, J., & Deng, J. (2015). HICO: A benchmark for recognizing human-object interactions in images. In Proceedings of the IEEE International Conference on Computer Vision. 1017–1025.

54. Carlson, A., et al. (2010). Toward an architecture for never-ending language learning. In AAAI.

55. Chemero, A. (2003). An outline of a theory of affordances. Ecological Psychology, 15(2), 181–

195.

56. Chen, D. L., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1. Association for Computational Linguistics, 190–200.

57. Chen, D. L., & Mooney, R. J. (2008). Learning to sportscast: A test of grounded language acquisition. In Proceedings of the 25th International Conference on Machine Learning. ACM, 128–135.

58. Chang, A. X., Savva, M., & Manning, C. D. (2014). Semantic parsing for text to 3d scene generation. In ACL 2014.

59. Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. L. (2015a). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

60. Chen, X., Shrivastava, A., & Gupta, A. (2013). Neil: Extracting visual knowledge from web data. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV). IEEE, 1409–1416.

61. Chen, Z., Lin, W., Chen, Q., Chen, X., Wei, S., Jiang, H., & Zhu, X. (2015b). Revisiting word embedding for contrasting meaning. In Proceedings of ACL.

62. Chelba, C., et al. (2014). One billion word benchmark for measuring progress in statistical language modeling. In Proceedings of the 15th Annual Conference of the International Speech Communication Association.

63. Chen, X., et al. (2015a). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

64. Chen, X., Shrivastava, A., & Gupta, A. (2013). Neil: Extracting visual knowledge from web data. In ICCV.

65. Chemero, A. (2003). An outline of a theory of affordances. Ecological Psychology, 15(2), 181–195.

66. Clark, S., & Pulman, S. (2007). Combining symbolic and distributional models of meaning. In AAAI Spring Symposium: Quantum Interaction. 52–55.

67. Cohen, M. D., & Bacdayan, P. (1994). Organizational routines are stored as procedural memory: Evidence from a laboratory study. Organiz. Sci., 5(4), 554–568.

68. Cohen, N., Sharir, O., & Shashua, A. (2016). On the expressive power of deep learning: A tensor analysis. In Proceedings of the 29th Annual Conference on Learning Theory. 698–728.

69. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. Syntax Sem. Struct. Stat. Transl., 103.

70. Choi, M. J., Torralba, A., & Willsky, A. S. (2012). Context models and out-of-context objects. Pattern Recogn. Lett., 33(7), 853–862.

71. Cohen, N., Sharir, O., & Shashua, A. (2016). On the expressive power of deep learning: A tensor analysis. In COLT.

72. Coradeschi, S., Loutfi, A., & Wrede, B. (2013). A short review of symbol grounding in robotic and intelligent systems. KI-Künstliche Intell., 27(2), 129–136.

73. Coradeschi, S., & Saffiotti, A. (2000). Anchoring symbols to sensor data: Preliminary report. In AAAI/IAAI. 129–135.

74. Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? Progr. Brain Res., 169, 323–338.

75. Darrell, T. (2010). Learning Representations for Real-world Recognition. UCB EECS Colloquium.

76. Das, P., Xu, C., Doell, R., & Corso, J. (2013). 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. 2634–2641.

77. Daumé III, H. (2007). Frustratingly easy domain adaptation. ACL 2007, 256.

78. Daumé III, H., Langford, J., & Marcu, D. (2009). Search-based structured prediction. Mach. Learn., 75(3), 297–325.

79. Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., & Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. In Advances in Neural Information Processing Systems. 1269–1277.

80. Dodge, J., et al. (2012). Detecting visual text. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 762–772.

81. Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2625–2634.