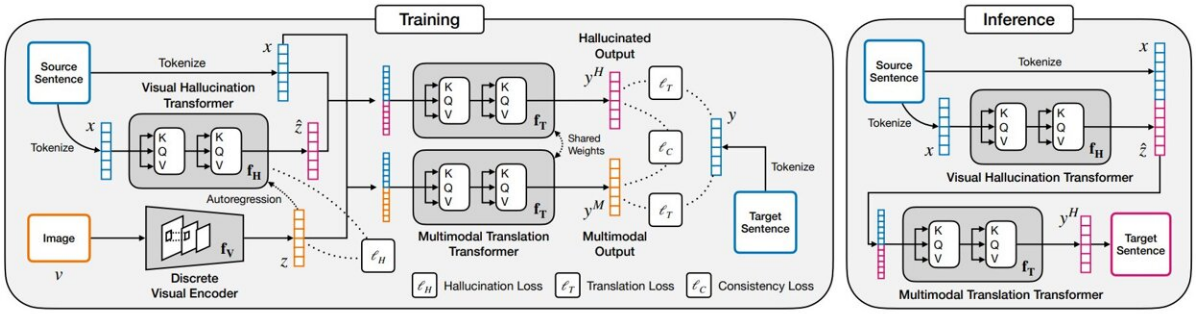
A machine-learning method hallucinates its way to better text translation

June 6 2022, by Lauren Hinkel



Overview of VALHALLA Architecture for Machine Translation. Left: Training pipeline of VALHALLA. Translation outputs are gathered from two streams of input, either with ground-truth visual tokens z or hallucinated representation zˆ, and optimized on a combination of hallucination, translation and consistency losses. Right: Inference process of VALHALLA in the absence of visual inputs. Credit: Li et al

As babies, we babble and imitate our way to learning languages. We don't start off reading raw text, which requires fundamental knowledge and understanding about the world, as well as the advanced ability to interpret and infer descriptions and relationships. Rather, humans begin our language journey slowly, by pointing and interacting with our environment, basing our words and perceiving their meaning through the context of the physical and social world. Eventually, we can craft full sentences to communicate complex ideas.

Similarly, when humans begin learning and translating into another language, the incorporation of other sensory information, like multimedia, paired with the new and unfamiliar words, like flashcards with images, improves language acquisition and retention. Then, with enough practice, humans can accurately translate new, unseen sentences in context without the accompanying media; however, imagining a picture based on the original text helps.

This is the basis of a new machine learning model, called VALHALLA, by researchers from MIT, IBM, and the University of California at San Diego, in which a trained neural network sees a source sentence in one language, hallucinates an image of what it looks like, and then uses both to translate into a target language. The team found that their method demonstrates improved accuracy of machine translation over text-only translation. Further, it provided an additional boost for cases with long sentences, under-resourced languages, and instances where part of the source sentence is inaccessible to the machine translator.

As a core task within the AI field of natural language processing (NLP), machine translation is an "eminently practical technology that's being used by millions of people every day," says study co-author Yoon Kim, assistant professor in MIT's Department of Electrical Engineering and Computer Science with affiliations in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and the MIT-IBM Watson AI Lab. With recent, significant advances in deep learning, "there's been an interesting development in how one might use non-text information—for example, images, audio, or other grounding information—to tackle practical tasks involving language," says Kim, because "when humans are performing language processing tasks, we're doing so within a grounded, situated world." The pairing of hallucinated images and text during inference, the team postulated, imitates that process, providing context for improved performance over current state-of-the-art techniques, which utilize text-only data.

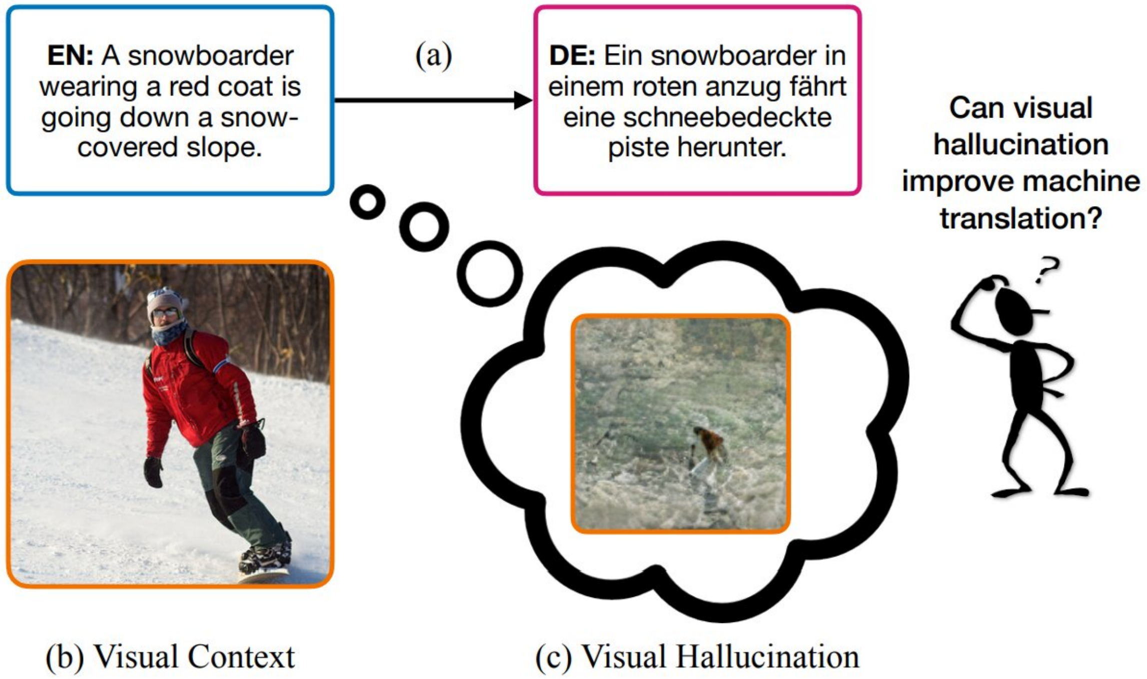
This research will be presented at the IEEE / CVF Computer Vision and Pattern Recognition Conference this month. Kim's co-authors are UC San Diego graduate student Yi Li and Professor Nuno Vasconcelos, along with research staff members Rameswar Panda, Chun-fu "Richard" Chen, Rogerio Feris, and IBM Director David Cox of IBM Research and the MIT-IBM Watson AI Lab.

Learning to hallucinate from images

When we learn new languages and to translate, we're often provided with examples and practice before venturing out on our own. The same is true for machine-translation systems; however, if images are used during training, these AI methods also require visual aids for testing, limiting their applicability, says Panda.

"In real-world scenarios, you might not have an image with respect to the source sentence. So, our motivation was basically: Instead of using an external image during inference as input, can we use visual hallucination—the ability to imagine visual scenes—to improve machine translation systems?" says Panda.

To do this, the team used an encoder-decoder architecture with two transformers, a type of neural network model that's suited for sequence-dependent data, like language, that can pay attention key words and semantics of a sentence. One transformer generates a visual hallucination, and the other performs multimodal translation using outputs from the first transformer.


EN: A snowboarder wearing a red coat is going down a snow-covered slope.
(a)
DE: Ein snowboarder in einem roten anzug fährt eine schneebedeckte piste herunter.
Can visual hallucination improve machine translation?
?
(b) Visual Context
(c) Visual Hallucination

Visual context such as images has been exploited in designing better machine translation systems. Different from most existing methods that require manually annotated sentence-image pairs as the input during inference, we introduce VALHALLA, that leverages hallucinated visual representation from the source sentences at test time for improved machine translation. Credit: Li et al

During training, there are two streams of translation: a source sentence and a ground-truth image that is paired with it, and the same source sentence that is visually hallucinated to make a text-image pair. First the ground-truth image and sentence are tokenized into representations that can be handled by transformers; for the case of the sentence, each word is a token. The source sentence is tokenized again, but this time passed through the visual hallucination transformer, outputting a hallucination, a discrete image representation of the sentence. The researchers incorporated an autoregression that compares the ground-truth and hallucinated representations for congruency—e.g., homonyms: a reference to an animal "bat" isn't hallucinated as a baseball bat. The hallucination transformer then uses the difference between them to optimize its predictions and visual output, making sure the context is consistent.

The two sets of tokens are then simultaneously passed through the multimodal translation transformer, each containing the sentence representation and either the hallucinated or ground-truth image. The tokenized text translation outputs are compared with the goal of being similar to each other and to the target sentence in another language. Any differences are then relayed back to the translation transformer for further optimization.

For testing, the ground-truth image stream drops off, since images likely wouldn't be available in everyday scenarios.

"To the best of our knowledge, we haven't seen any work which actually uses a hallucination transformer jointly with a multimodal translation system to improve machine translation performance," says Panda.

Visualizing the target text

To test their method, the team put VALHALLA up against other state-of-the-art multimodal and text-only translation methods. They used public benchmark datasets containing ground-truth images with source sentences, and a dataset for translating text-only news articles. The researchers measured its performance over 13 tasks, ranging from translation on well-resourced languages (like English, German, and French), under-resourced languages (like English to Romanian) and non-English (like Spanish to French). The group also tested varying transformer model sizes, how accuracy changes with the sentence length, and translation under limited textual context, where portions of the text were hidden from the machine translators.

The team observed significant improvements over text-only translation methods, improving data efficiency, and that smaller models performed better than the larger base model. As sentences became longer, VALHALLA's performance over other methods grew, which the researchers attributed to the addition of more ambiguous words. In cases where part of the sentence was masked, VALHALLA could recover and translate the original text, which the team found surprising.

Further unexpected findings arose: "Where there weren't as many training [image and] text pairs, [like for under-resourced languages], improvements were more significant, which indicates that grounding in images helps in low-data regimes," says Kim. "Another thing that was quite surprising to me was this improved performance, even on types of text that aren't necessarily easily connectable to images. For example, maybe it's not so surprising if this helps in translating visually salient sentences, like the 'there is a red car in front of the house." [However], even in text-only [news article] domains, the approach was able to improve upon text-only systems."

While VALHALLA performs well, the researchers note that it does have limitations, requiring pairs of sentences to be annotated with an image, which could make it more expensive to obtain. It also performs better in its ground domain and not the text-only news articles. Moreover, Kim and Panda note, a technique like VALHALLA is still a black box, with the assumption that hallucinated images are providing helpful information, and the team plans to investigate what and how the model is learning in order to validate their methods.

In the future, the team plans to explore other means of improving translation. "Here, we only focus on images, but there are other types of a multimodal information—for example, speech, video or touch, or other sensory modalities," says Panda. "We believe such multimodal grounding can lead to even more efficient machine translation models, potentially benefiting translation across many low-resource languages spoken in the world."

More information: VALHALLA: Visual Hallucination for Machine Translation

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: A machine-learning method hallucinates its way to better text translation (2022, June 6) retrieved 9 April 2024 from https://techxplore.com/news/2022-06-machine-learning-method-hallucinates-text.html