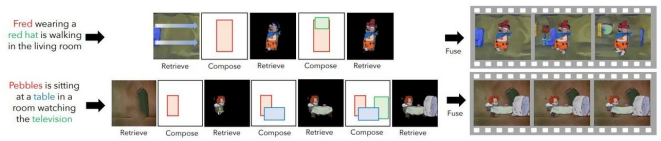
Researchers explore working up cartoons using text descriptions

24 April 2018, by Nancy Owano



Given a novel description, Craft sequentially composes a scene layout and retrieves entities from a video database to create complex scene videos. Credit: arXiv:1804.03608 [cs.CV]

What if you were told you can create cartoons by just working off text descriptions?

Reports are in that a group of researchers unveiled an AI capable of making original videos of "The Flintstones" from text descriptions.

Yes, these are scenes created by an artificial intelligence. Consider some scene description: Fred is wearing a blue hat and talking to Wilma in the living room. Wilma then sits down on a couch.

Composition, Retrieval and Fusion Network, or CRAFT, is the name of their model. The authors noted they showed CRAFT on Flintstones, a dataset with over 2,500 videos and each 75 frames long.

They have written a paper, titled "Imagine This! Scripts to Compositions to Videos" and it is on arXiv. The five researchers are Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem and Aniruddha Kembhavi. Author affiliations include The Allen Institute for Artificial Intelligence (AI2), The University of Illinois Urbana-Champaign and The University of Washington.

The authors said that once it is given a novel description, "Craft sequentially composes a scene layout and retrieves entities from a video database to create complex scene videos."

Tristan Greene, *The Next Web*, explained how the technology works: "Craft uses the annotations from videos to determine how the original images correspond to the words used to describe them. Eventually it builds up a set of parameters that enables it to 'understand' what makes individual characters and objects from the cartoon match their plain-language counterparts. Once it understands this relation, it's able to generate video clips based on novel text inputs that look a lot like the cartoon it was trained on."

The authors also discussed their model based on text:

"Unlike pixel generation approaches, our appearance model is based on text to entity segment retrieval from a video database. Spatio-temporal segments are extracted from the retrieved videos and fused together to generate the final video. The layout composition and entity retrieval work in a sequential manner which is determined by the language input."

The authors stated that "CRAFT outperforms direct pixel generation approaches."

Interestingly, video viewers wrote responses ranging from wow to tepid to confused.

Several thought it was Awesome; one remarked that it was "more advanced than I would have imagined" and another said "it still looks like if someone tried to animate for the first time on demo software. It looks like it has potential, though."

Another observer was more confused than startled. "I'm confused. My understanding is that the AI learned 25k fully annotated cartoons. And then the researchers typed in a text scenario, and the AI just found images that matched it? Isn't that just a simple retrieval of the corresponding video snippet

based on a text lookup from the annotated database? What am I missing?"

Writers on tech sites offered their perspective about this research. Referring to the videos, *The Next Web* stepped in. OK it's a "glitchy little clip," as Tristan Greene put it. All the same, he added, "Today's glitchy little clip, generated from simple text phrases, could lead to tomorrow's entertainment being created from scratch by AI instead of studios full of people."

Andrew Liszewski in *Gizmodo* similarly found that the quality of the animations that were generated was "awful at best" and "no one's going to be fooled into thinking these are the Hanna-Barbera originals." Nonetheless, he added, seeing an AI generate a cartoon, with iconic characters, all by itself, was "a fascinating sneak peek at how some films and TV shows might be made one day."

Lucy Black wrote Sunday, in *I Programmer* that "This is more than just another clever trick with neural networks. It is a sign that AI is moving towards larger systems where deep neural networks do different jobs and work together to create the solution. You could call it the second stage of deep neural networks."

OK, unanswered question: Would animators lose their jobs. Black said, "Yes I suppose given time and effort something like CRAFT could be developed into a cartoon generator and throw thousands of animators out of a job, but computer graphics is already chipping away at that job market."

More information: Imagine This! Scripts to Compositions to Videos, arXiv:1804.03608 [cs.CV] arxiv.org/abs/1804.03608

Abstract
Imagining a scene described in natural language with realistic layout and appearance of entities is the ultimate test of spatial, visual, and semantic world knowledge. Towards this goal, we present the Composition, Retrieval, and Fusion Network (CRAFT), a model capable of learning this knowledge from video-caption data and applying it while generating videos from novel captions. CRAFT explicitly predicts a temporal-layout of mentioned entities (characters and objects), retrieves spatio-temporal entity segments from a video database and fuses them to generate scene videos. Our contributions include sequential training of components of CRAFT while jointly modeling layout and appearances, and losses that encourage learning compositional representations for retrieval. We evaluate CRAFT on semantic fidelity to caption, composition consistency, and visual quality. CRAFT outperforms direct pixel generation approaches and generalizes well to unseen captions and to unseen video databases with no text annotations. We demonstrate CRAFT on FLINTSTONES, a new richly annotated video-caption dataset with over 25000 videos.