Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CREATING INFORMATION-MAXIMIZING NATURAL LANGUAGE MESSAGES THROUGH IMAGE CAPTIONING-RETRIEVAL

FABIAN KARL, MIKKO LAURI & CHRIS BIEMANN AT KONVENS 2019

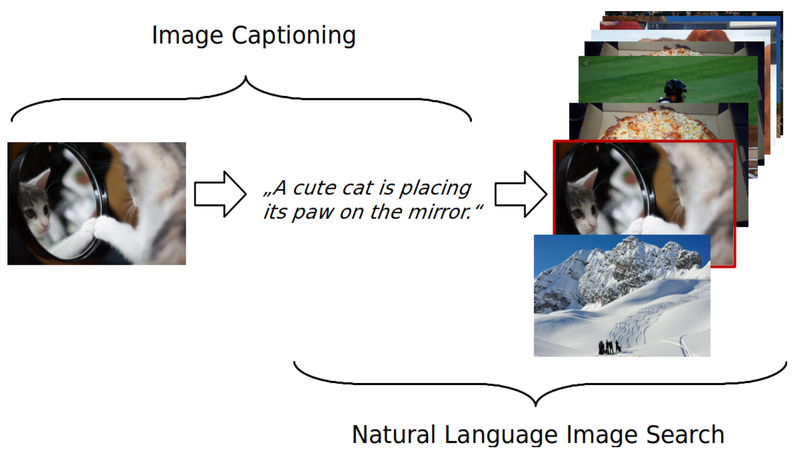
ABSTRACT

We propose the ICR problem that casts the objective of language generation as information exchange. To solve the ICR problem, we design and implement an end-to-end neural network architecture that **describes the content of images in natural language, and retrieves them solely based on these generated descriptions**. The main goal is to be able to generate **information-maximizing natural language messages**. We experimentally show a **strong increase in message information content** while losing some grammatical correctness in the generated descriptions in a **semi-supervised setting** where caption generation is trained towards retrieval quality.

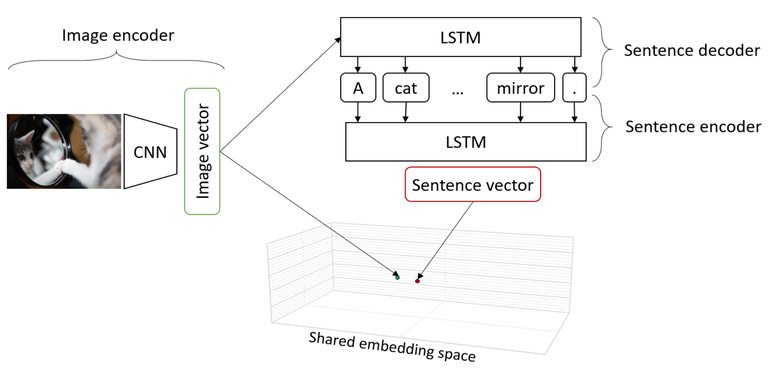
Image Captioning (IC):
Automatically describing the content of a given image.

Natural Language Image Search (NLIS):
Retrieving an image based on a given description.

Image Captioning-Retrieval (ICR) :
An image is first automatically annotated and then retrieved among a number of candidate images.



IMPLEMENTATION



Our ICR model above is a combination of an IC network (image encoder and sentence decoder) and an NLIS model (image encoder and sentence encoder).

The IC model transforms an image into a natural language description of its content. It is implemented by a state-of-the-art CNN-LSTM encoder-decoder architecture.

Our NLIS model is implemented though the same image encoder and an additional sentence encoder. The encoded feature representations of image and sentence are mapped in a shared latent space, where the distance between image-sentence pairs is interpreted as similarity.

DATA

The Microsoft Common Objects in Context (MSCOCO) dataset with 2017 split was used for training, validation and testing in all our experiments, containing 118,287 training and 5,000 validation images with five annotations each.

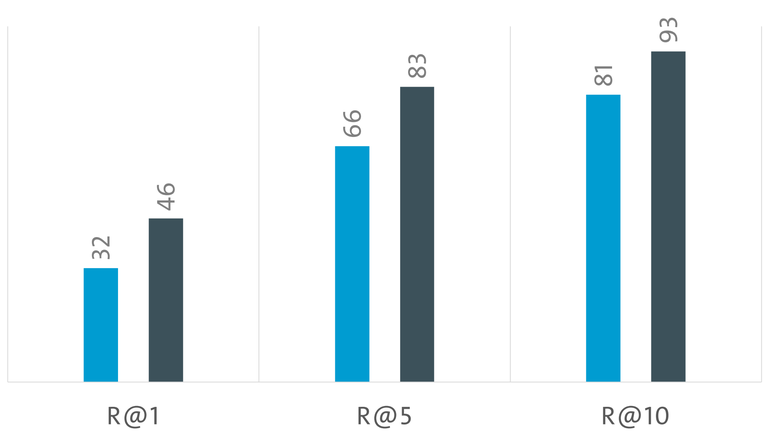
TRAINING SETUP

1. Both IC and NLIS model are respectivel trained on their own task until convergence.
2. The weights are transferred to our ICR network.
3. The ICR network is finetuned on a combination of triplet ranking loss from the retrieved images and the cross-entropy between the generated and the ground truth annotations.

RESULTS

The chart below shows the image retrieval scores for our best NLIS model trained on ground truth annotations from MSCOCO compared to the best scores from our best ICR model.

The score shows the average percentage of images retrieved within the top 1, 5 or 10 ranks. When we let the model create and retrieve descriptions, the amount of transferred information can be increased greatly. This is visible by a higher image retrieval score.



DISCUSSION & CONCLUSION

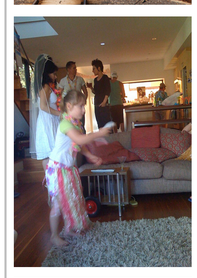
We show how training an IC network with a more implicit objective can improve the amount of information captured in generated descriptions. The newly generated sentences are not grammatically perfect but understandable by humans. They often capture more details and describe aspects of the images that are not even presented in the ground truth data.

This shows how our objective trains a system towards transferring information, while still creating human-readable sentences.

a train is parked on the tracks in a city .

A train traveling past tall white buildings on train tracks.
A locomotive train carrying carts down a track.
A train is coming down the tracks near a building.

an old locomotive train caboose on an railroad train tracks wires wires wires wires overpass overpass



a group of people standing in a room with a large screen .

A young girl holding a controller playing a video game.
A young girl playing a video game while others talk.
A little girl holding a white Nintendo Wii game controller.

three men boys women standing on a couch couch gifts gifts gifts living room living room

a man is playing tennis on a court .

Two men shaking hands while standing on a tennis court.
Two tennis players shaking hands over the net
Two tennis players are facing each other over a tennis net.

two men man standing on an tennis court on an tennis court net courts courts courts

a man riding on the back of an elephant .

A man riding on the back of a tusked elephant by a muddy river
A man riding on the back of an elephant along a dirt road.
Man riding an elephant up a hill near a field.

two man riding on elephant elephant walking through some river river saddle a river stream river

IC single model ground truth ICR combined model