Deep Neural Networks (e.g., VGG-19, Simonyan et al., 2014) trained for object classification using the large ImageNet database (organized along the nouns of the WordNet hierarchy) develop feature detectors similar to those in the human visual system. DeepGaze II (Kümmerer et al., 2016) uses VGG19 feature responses to train a second Neural network to predict empirical saliency, i.e. pseudo-fixations from deep networks trained on photographs of real-world scenes, with quite promising first results.

Subjective Description

The foundation of this endeavor is a large collection of graphic novels, which are annotated using a variety of methods. Part of the annotations are provided by human coders, and include high-level meta data about the works, and mid-level descriptions of pages and panels, including the locations and identities of actors / characters, text, objects, and panel transitions. To aid human coders, we developed an annotation tool and a corresponding XML dialect, extending CBML (Walsh, 2012).

Objective Description

Since human descriptions necessarily provide a degree of subjectiveness, we also include more objective lower-level descriptions of visual elements in terms of feature descriptors developed in computer vision. The DeepGaze II predictions (blue dots) and DeepGaze-Il predictions (red contours) are in surprisingly good agreement.

Annotations are augmented by eye-tracking data for a subset of pages, reflecting where readers direct their attention (empirical saliency). The corpus can therefore also be used for evaluating computational models of saliency (theoretical saliency).

Deep Neural Networks: VGG-19 + DeepGaze II

Applications

An annotated corpus of graphic novels is a valuable resource for linguistics and literature studies. For example, stylometric analyses of text, but also of visual features becomes possible. The set of tools and methods is potentially useful for other fields, such as arts history (cf. Elgammal & Saleh, 2015). Availability of ground truth annotations provides training material for object recognition and classification tasks in computer science and data science. Cognitive psychology might be interested in using the combination of eye movement data, image and text descriptions for evaluating models of attention during reading and scene perception. The AOI-based annotations additionally allow for interesting experimental tests of information integration across panels or narrative construction. More generally, a set of tools based on image processing might benefit various areas of the Digital Humanities.

Improving the annotation tool

Panels segmentation

Automatically locating text in a comics panel is surprisingly difficult, given the ease with which humans can perform the task. While classical algorithms are satisfactory for panel detection, they perform less well for recognition of other elements such as headings, speech bubbles, captions, or persons. Deep features might be useful for these tasks. For example, some neurons in the upper convolutional layers of pre-trained VGG-19 can be used as a speech bubble detectors. Additional training using gold standard annotations is likely to yet improve the results. Projected further developments include recognition of faces and protagonist identities.

Text spotting

Annotation is a somewhat tedious and time-consuming process for humans. Image analysis can be used to alleviate some of the steps. For example, segmentation of the page into panels is fairly straightforward and works well in many cases, using a combination of flood-fill and connected-components operation, given some initial heuristic guess (or human input) of which is the likely background color, and some post-processing discarding very small regions.

Automaticing the addition of high-level notations additionally allows for interesting experimental tests of information integration across panels or narrative construction. More generally, a set of tools based on image processing might benefit various areas of the Digital Humanities.

References


HOG SIFT GIST