Adverserial Examples

Jan Plank,[*] Philipp Höhne,[†] and Jonas Dehning[‡]
Universität Göttingen
(Dated: July 11, 2017)

In this work we generate adversarial examples for convolutional networks trained on two different datasets. Two different methods are compared for the generation: the first is to add a small "noise" in the direction of the gradient of the loss function, the other is to minimize a custom function which allows to find an example nearer to the original image than with the gradient method. Than the robustness of the network to adversarial examples are compared as function of the depth of networks, for the two datasets.

I. INTRODUCTION

Neural networks

A. Convolutional Neural Networks

Convolutional neural networks are used to categorize images. The are composed of successive convolutional and pooling layers. In the convolutional layer each pixel (neuron) gets the input of the surrounding pixels of the previous layer or the image, and multiply it by a matrix, thus making a convolution. The convolutional matrix is the same for the whole image but typically for each convolutional layer a set of matrices each one applied to the whole image, also called filters, are used.

1. *Wide text (A level-3 head)*

The `widetext` environment will make the text the width of the full page, as on page **??**. (Note the use the `\pageref{#1}` command to refer to the page number.)

a. Note (Fourth-level head is run in) The width-changing commands only take effect in two-column formatting. There is no effect if text is in a single column.

B. Citations and References

A citation in text uses the command `\cite{#1}` or `\onlinecite{#1}` and refers to an entry in the bibliography. An entry in the bibliography is a reference to another document.

1. *Citations*

Because REVTEX uses the `natbib` package of Patrick Daly, the entire repertoire of commands in that package are available for your document; see the `natbib` documentation for further details. Please note that REVTEX requires version 8.31a or later of `natbib`.

a. Syntax The argument of `\cite` may be a single *key*, or may consist of a comma-separated list of keys. The citation *key* may contain letters, numbers, the dash (-) character, or the period (.) character. New with natbib 8.3 is an extension to the syntax that allows for a star (*) form and two optional arguments on the citation key itself. The syntax of the `\cite` command is thus (informally stated)

`\cite {` *key* `}`, or
`\cite {` *optarg+key* `}`, or
`\cite {` *optarg+key* `,` *optarg+key...* `}`,
where *optarg+key* signifies

key, or
`*`*key*, or
`[`*pre*`]`*key*, or
`[`*pre*`][`*post*`]`*key*, or even
`*[`*pre*`][`*post*`]`*key*.
where *pre* and *post* is whatever text you wish to place at the beginning and end, respectively, of the bibliographic reference (see Ref. [**?**] and the two under Ref. [**?**]). (Keep in mind that no automatic space or punctuation is applied.) It is highly recommended that you put the entire *pre* or *post* portion within its own set of braces, for example: `\cite { [ {`*text*`}]`*key*`}`. The extra set of braces will keep LATEX out of trouble if your *text* contains the comma (,) character.

The star (*) modifier to the *key* signifies that the reference is to be merged with the previous reference into a single bibliographic entry, a common idiom in APS and AIP articles (see below, Ref. [**?**]). When references are merged in this way, they are separated by a semicolon instead of the period (full stop) that would otherwise appear.

b. Eliding repeated information When a reference is merged, some of its fields may be elided: for example, when the author matches that of the previous reference, it is omitted. If both author and journal match, both are omitted. If the journal matches, but the author does not, the journal is replaced by *ibid.*, as exemplified by Ref. [**?**]. These rules embody common editorial practice in APS and AIP journals and will only be in effect if the markup features of the APS and AIP BibTEX styles is employed.

[*] janhendrik.plank@stud.uni-goettingen.de
[†] philipp.hoehne@stud.uni-goettingen.de
[‡] j.dehning@stud.uni-goettingen.de

c. The options of the cite command itself Please note that optional arguments to the *key* change the reference in the bibliography, not the citation in the body of the document. For the latter, use the optional arguments of the `\cite` command itself: `\cite *[`*pre-cite*`]` `[`*post-cite*`]{`*key-list*`}`.