MLCV2017: Multi-Layer Knowledge Transfer for Neural Networks

Lennard Kiehl
lennard.kiehl@gmail.com

Roman Remme
roman.remme@gmx.de

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

The idea of transferring knowledge between different architectures of neural networks, specifically from bigger models to smaller models, has been introduced in [6]. Part of the motivation for this process called distilling is to create a smaller model, which is faster at runtime, with the same knowledge as the bigger model. Distilling works by introducing an additional term to the loss function that links the last layers (logits) of both networks by calculating the cross entropy between them. It was shown in [6] that this alone serves as a very good knowledge transfer tool. We want to extend on this idea and also add links between intermediate layers. The popular VGG-16 model introduced in [8] serves as the bigger model and the goal is to distill each group of convolutional layers into only one convolutional layer, for an overview of the architectures see Table 1.1. We investigate how training hyperparameters influence the process of successfully distilling knowledge.

1.1. Models

As a big model to distill knowledge from we use VGG-16 (see [8]). For the smaller model that was trained with the help of the big one a similar architecture was used, where the number of convolutional layers between pooling layers was reduced from to just one. We also cut one of the fully connected layers (see Table 1.1). The similarity made it possible to compare intermediate activations at many points in the model.

Network Configurations with linkable layers		
	VGG-16	VGG-7
	input (224×224 RGB image)	
link 1	conv3-64 conv3-64	conv3-64
	maxpool 2×2	
link 2	conv3-128 conv3-128	conv3-128
	maxpool 2×2	
link 3	conv3-256 conv3-256 conv3-256	conv3-256
	maxpool 2×2	
link 4	conv3-512 conv3-512 conv3-512	conv3-512
	maxpool 2×2	
link 5	conv3-512 conv3-512 conv3-512	conv3-512
	maxpool 2×2	
link 6	FC-4096 FC-4096 FC-10	FC-4096 FC-10
	softmax	

Table 1. **Network configurations.** The convolutional layers are denoted as conv(*kernel size*)-(*number of channels)* and the fully connected layers as FC-(*number of output channels*). ReLu units are omitted for brevity. The leftmost column gives the links between both networks that are added to the loss function.

1.2. Loss

The loss function consists of three terms:

- The "hard" loss: The cross-entropy of the output distribution of the network with the correct labels
- The "soft" last layer loss: The cross-entropy of the output distribution with the "soft-targets", the output of the cumbersome model to extract knowledge from. Here, softmax layers with temperature T are used.

Temperature	Test set accuracy
0.6	76.3 %
1	76.5 %
1.5	77.0 %
2	**77.4** %
2.5	76.7 %
3	77.2 %
5	73.1 %
10	64.4 %
40	67.3 %

Table 2. **Last layer transfer results.** bla bla

Linked layers	β	Test set accuracy
3	10	85.9%
2, 3, 4, 5	10	87.0%
2, 3, 4, 5	40	87.9%
5	10	87.7%
3, 4	10	84.8%
1, 2, 3, 4, 5, 6	10	81.2%
2, 3, 4, 5, 6	10	78.5%

Table 3. **Intermediate Layers transfer results.** used last layer with temperature 2 and $\alpha = 10$

- The intermediate layer loss: This is the mean squared error between the activations of pairs of layers in a certain set.

The third term is the new part of our approach. A theoretical advantage is that training times should be reduced, as gradients do not have to be propagated through the whole network to reach the first layers. Also this is a

1.3. Dataset

We use CIFAR-10 [7] as the dataset to train to train both models on. It consists of 50000 training and 10000 test RGB images of size 32×32 pixels. Each image belongs to one of ten classes. To use the standard VGG architecture with these low-resolution images, they are scaled up to 224×224 pixels. Each image is preprocessed by subtracting the mean RGB value, computed on the training set, from each pixel.

2. Training

We used stochastic gradient descent with momentum 0.9 (CITATION NEEDED?) as an optimizer. We started with a learning rate of 0.004 and let it decay by a factor of 10 every 10 epochs for 25 epochs.

3. Results

3.1. Blind review

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one's own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

Blind review means that you do not use the words "my" or "our" when citing previous work. That is all. (But see below for techreports.)

Saying "this builds on the work of Lucy Smith [1]" does not say that you are Lucy Smith; it says that you are building on her work. If you are Smith and Jones, do not say "as we show in [7]", say "as Smith and Jones show in [7]" and at the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper just asking to be rejected:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of our previous paper [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Removed for blind review

An example of an acceptable paper:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of the paper of Smith *et al.* [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Smith, L and Jones, C. "The frobnicatable foo filter, a fundamental contribution to human knowledge". Nature 381(12), 1-213.

If you are making a submission to another conference at the same time, which covers similar or overlapping material, you may need to refer to that submission in order to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission [4] as additional material and cite it as

> [1] Authors. "The frobnicatable foo filter", F&G 2014 Submission ID 324, Supplied as additional material `fg324.pdf`.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a techreport for further details. Thus, you may say in the body of the paper “further details may be found in [5]”. Then submit the techreport as additional material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool which is widely known to be restricted to a single institution. For example, let’s say it’s 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled “Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties”, by Zeus *et al*.

You can handle this paper like any other. Don’t write “We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]”. That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] didn’t handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don’t you know. It displayed the following behaviours which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ: Are acknowledgements OK? No. Leave them for the final copy.

3.2. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TeXbook, p165.

The space after *e.g*., meaning “for example”, should not be a sentence-ending space. So *e.g*. is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

When citing a multi-author paper, you may save space by using “et alia”, shortened to “*et al*.” (not “*et. al.*” as “*et*” is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: “ Frobnication has been trendy lately. It was introduced by Alpher [1], and subsequently developed by Alpher and Fotheringham-Smythe [2], and Alpher *et al*. [3].”

This is incorrect: “... subsequently developed by Alpher *et al*. [2] ...” because reference [2] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*.

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [2, 1, 4] to [1, 2, 4].

4. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5 × 11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

4.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high. Page numbers should be in footer with page numbers, centered and .75 inches from the bottom of the page and make it start at the correct page number rather than the 4321 in the example. To do this fine the line (around line 23)

```
%\ifcvprfinal\pagestyle{empty}\fi
```

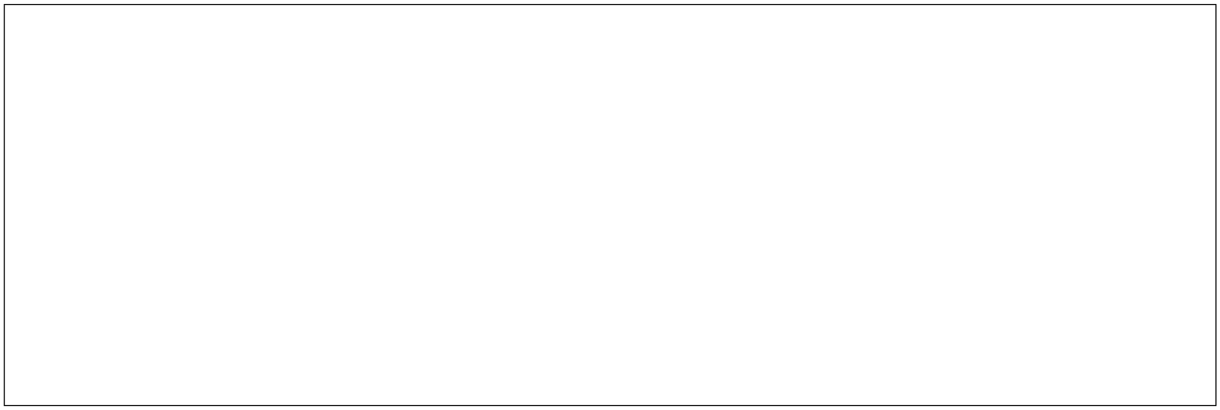

Figure 2. Example of a short caption, which should be centered.

```
\setcounter{page}{4321}
```

where the number 4321 is your assigned starting page.

Make sure the first page is numbered by commenting out the first page being empty on line 46

```
%\thispagestyle{empty}
```

4.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures 1 and 2. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 4. Results. Ours is better.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

4.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

4.4. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [4]. Where appropriate, include the name(s) of editors of referenced books.

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

4.5. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

4.6. Color

Please refer to the author guidelines on the CVPR 2017 web page for a discussion of the use of color in your document.

5. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

References

[1] A. Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002.

[2] A. Alpher and J. P. N. Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003.

[3] A. Alpher, J. P. N. Fotheringham-Smythe, and G. Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004.

[4] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`.

[5] Authors. Frobnication tutorial, 2014. Supplied as additional material `tr.pdf`.

[6] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[7] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.