中国科学院大学

University of Chinese Academy of Sciences

文章写作规范研究报告

第 X 小组

小组组长：XXX

小组成员：XXX XXX XXX XXX

2020 年 12 月

第 1 章　关于本报告的说明

本报告是针对是提出神经网络经典模型 AlexNet2012 的论文《ImageNet Classification with Deep Convolutional Neural Networks》进行的写作规范研究报告。

1.1　组织架构

本报告的主体部分为第 2-5 章。

其中主体的第一部分是第 2章，分析的是该文章的标题、署名、摘要和引证等，并对摘要的结构和内容提出了一些值得商榷的意见。

本报告的主体第二部分是第 3章、第 4章，分析的是该文章的主体结构，其中第 3章是对文章引言的具体分析，第 4章是对文章方法、结果和讨论部分的具体分析。

本报告的主体第三部分是第 5章，分析该文章中的图表风格特点，并提出了一些改进意见。

本报告的头尾两小章是对本报告的说明和小结。

1.2　成员分工

本小组的成员分工如下表所示：

表 1.1 小组成员分工

成员分工	负责的部分	对应的正文内容
XXX	分析文章中的方法、结果和讨论	第 4章
XXX	分析文章中的图表	第 5章
XXX	分析文章的引言	第 3章
XXX	分析论文的标题、署名、摘要、引证	第 2章
XXX	整合、报告排版	头尾两小章

第 2 章　标题、摘要、引证、署名

2.1　标题

文章的标题（如图 2.1）简洁有力，以一句话文摘的形式反映了该文章研究的任务以及采用的方法这两项文章的精髓。标题的内容非常具体、准确、简练，均由术语构成，每个词都有明确的含义。在用词上，均采用名词短语，语法规范。

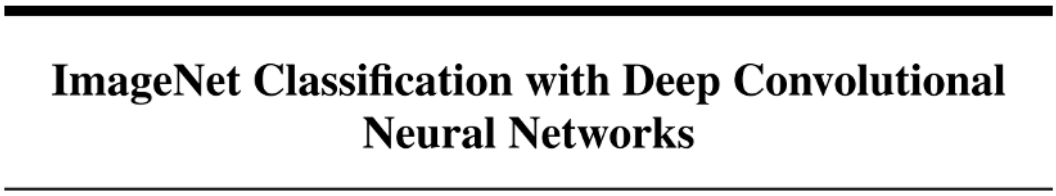


图 2.1 文章的标题

2.2　署名

在署名的设计（如图 2.2）上，作者名称、单位以及邮箱采用不同的字体字号进行区分，并且排版工整，使得读者对于作者的信息一目了然。在署名的顺序上，可以看出，并不是按严格按名称首字母进行的排序，经过对作者的背景的研究，发现前两位作者是最后一位作者 Hinton 的学生，则署名顺序是以学生-导师的顺序完成，学生部分的署名按名称首字母的顺序排列。需要说明的是，由于该文章的三位作者都是深度学习领域公认的明星学者，因此该文章或许不存在所谓需要特别突出的“第一作者”。

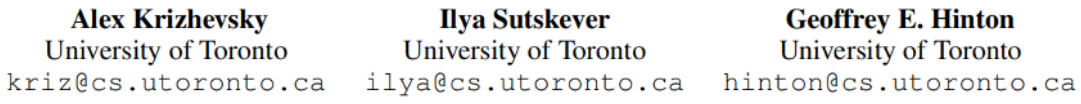


图 2.2 文章的署名

2.3　摘要

文章摘要的结构分析结果如图 2.3所示。

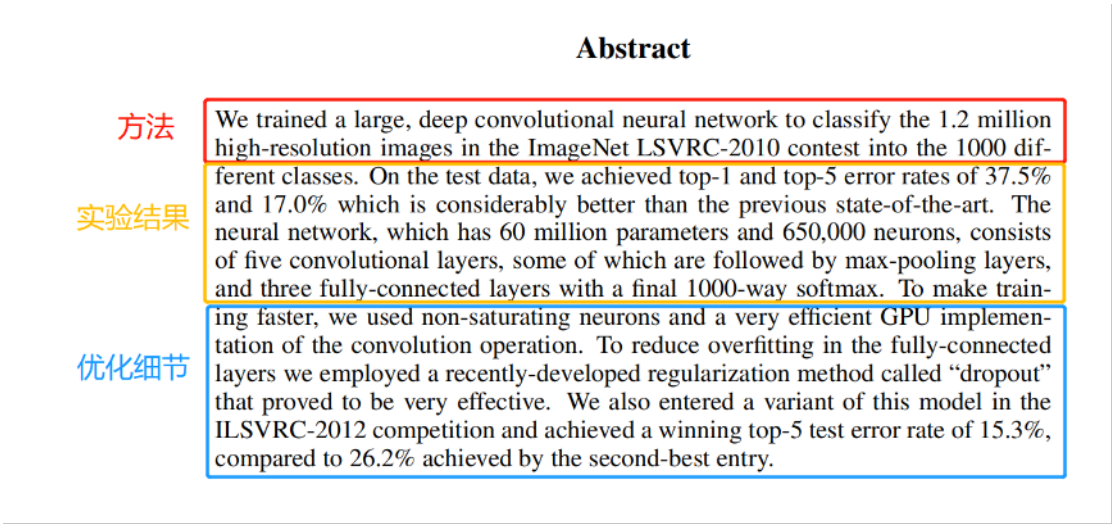


图 2.3 文章的摘要

在结构上，摘要采取了叙述文章采用的主要方法，即使用深度卷积神经网络进行图片分类，随后对实验结果进行了描述，最后通过叙述优化细节，来说明该文的方法是如何取得更优的效果。

这种摘要的结构是值得商榷的，一方面，对于大同行以及外行来说，这篇摘要并不完美，摘要并没有给出文章所面向的问题和研究意义，通过摘要无法对文章所面向的问题进行一个初步了解，如果将来论文需要面向大同行，建议将问题的背景进行补充；另一方面，对于小同行来说，这篇摘要已经能够充分反映研究中的必要信息，对于该文章所在领域来说，提高分类效率这一问题已经是此类与新模型相关的论文的共识性问题，关键在于新模型的结构以及效果如何，因此通过阅读这一摘要，即可清晰了解文章的工作及其意义。

在摘要内容的撰写方面，采用了简洁的纯文本进行叙述，语法上也没有较大的瑕疵，是值得肯定的。但摘要内容中出现了许多细节性数据，这一点是值得商榷的。摘要中给出了大量的细节性数据用于描述神经网络的模型架构、数据规模以及实验结果，显得不够简明扼要，如果将来论文需要面对大同行，建议将这些具体的实验数据略去，采用高度概括的词语进行叙述。但另一方面，对于小同行来说，根据这些细节性数据，已经可以较为全面的了解文章的工作以及效果。比如文章涉及领域面向大规模的图片分类问题，对于大同行来说，摘要中采用“大规模”这一概念描述即可，但对小同行来说，采用具体的数字规模说话，如该文中的“1.2 million”，更能直接反映文章模型的优势。此外，小同行还能够根据摘要中给出的一些模型关键参数的具体数值对文章设计的模型进行复现。

2.4 引证

对于引证部分，该文章均采用如图 2.4所示的间接引用方式，引用的格式较为规范，参考文献的序号均紧跟在引证的主要内容后，且对于所引证的内容都通过简练的语言进行了转述，使之符合文章的表达情境。在文内的引用没有太大的问题，加之该文章作为深度学习领域的代表性论文，作者又是业界著名学者，其引证方式值得参考学习。

necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

图 2.4 文章中的引证示例

对于引证的文献的范围来说，如图 2.5，可以看到，该文的引证范围较为全面，且参考文献的质量都较高。一方面，其覆盖了涉及领域早期的经典的奠基性文献，以及论文发表时期的最新研究成果；另一方面，其引用论文的质量较高，可以看到，引用文献许多来自于文章所属行业的顶级会议和顶级作者。此外，该文的引证文献均为公开文献，使得文章的引证是确实有据可查的。

[1] R.M. Bell and Y. Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.
[2] A. Berg, J. Deng, and L. Fei-Fei. Large scale visual recognition challenge 2010. www.image-net.org/challenges. 2010.
[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
[4] D. Cireşan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. *Arxiv preprint arXiv:1202.2745*, 2012.
[5] D.C. Cireşan, U. Meier, J. Masci, L.M. Gambardella, and J. Schmidhuber. High-performance neural networks for visual object classification. *Arxiv preprint arXiv:1102.0183*, 2011.
[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
[7] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. *ILSVRC-2012*, 2012. URL http://www.image-net.org/challenges/LSVRC/2012/.
[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

代表性优质论文
经典著作
文章发表时期的最新成果

图 2.5 文章的参考文献

第 3 章　引言部分的分析

在 Introduction 部分，作者完成了五部分内容的说明，分别对应文章中的五段。

第一段，作者对图像分类这一领域简单的做了背景和现状的介绍，这一段内容相信不只是小同行，大同行都能轻松读懂。

第二段作者主要说明了如果在大规模数据集下使用前人方法会遇到的问题和挑战——任务的复杂性会非常高。

第三段作者根据上一段的问题和挑战提出了解决思路，Gpu 和 2D 卷积相结合可以训练大型的 CNN，且带标签数据集可以使模型不发生严重的过拟合。

第四段作者通过和前人方法的效果对比说明论文的贡献，在说明使用方法的同时，说明了文章的具体结构。

最后，在引言第五段，作者提出了可能存在的改进方法——只需要等更快的 GPU 和更大的数据集就可以得到更好的结果。

这五个部分内容同刘老师上课讲的完全一致，拥有了目的、核心任务、思路、贡献和文章结构 5 个要素。通过阅读引言，就能明白了作者想要解决什么样的问题，且该文作者只着重讲了一个问题的解决，也就是只有一个卖点——在大规模图片数据集上进行对象识别得到了比其他方法更好的结果，而这一点也是刘老师上课时强调的。

值得一提的是，该文提出问题的方式是“补”，也就是提出了 Gap，在前人工作的基础上进行了改良，得到了比前人更好的结果。

纵观整个引言部分，作者精雕细琢，没有多写一句废话，没有介绍常识，也没有提及具体细节，对他人的工作也非常客观的做出了评价。

第 4 章　方法、结果、讨论

接下来，我们将对该文章使用的方法、得出的结果，以及该文章的讨论部分进行写作规范方面的分析。

4.1　方法

这篇文章在引言部分，就点出了当前三大主流研究方向——更大的数据集，具有更强大学习能力的模型，更好地处理过拟合的方法。而这三点也是一个神经模型实现的关键。方法部分，作者也是这样组织的。

首先，作者在第二部分数据集中，对模型采用的训练集和测试集进行了详细地介绍——ImageNet。由于本篇文章重点突出，在较大数据集上，如何训练出具有强大学习能力的模型，所以作者在数据集介绍过程中，突出强调 ImageNet 是一个超大型的高分辨率图像数据集。同时，作者说明了他们在训练过程中观测的重要指标，错误率 (top-1 和 top-5)。另外，作者也描述了他们对数据的预处理，为模型效果复现提供了必要的基础。

方法第三部分便是模型的框架。作者开篇介绍网络共 8 层——5 个卷积层和 3 个全连接层，这好比是一个具体模型的骨架。在 3.1 至 3.4 节中，作者针对模型架构中新颖的特点，按照重要程度，一一进行了介绍。这些好比是依附在骨架上的血肉，而正是这些与前人不同的特点，使得 Alexnet2012 成为了当时图像分类最好的模型。其中包括非线性 ReLU 方法，更是成为了今后解决梯度消失问题经典方法。

方法的第四部分则是针对处理过拟合问题的方法，结构与第三部分相似，分条列出两种主要方法——数据增强和 Dorpout。

4.2　结果

文章的结果采用报告实验型结果，如表 5.1、表 5.2所示。文章给出了 Alexnet 模型与其他人（表中的斜体部分）效果最好的模型的 Top-1 和 Top-5 错误率，清晰直观地显现出 Alexnet 模型在图像分类上的优越性。同时，在定量描述之外，作者还对结果进行了定性评估，分析模型在特定数据集上的效果，图文并茂，直

观明了。

4.3　讨论

作者在讨论部分对全文进行了总结，在强调模型在具有挑战性的大型数据集上取得破纪录的结果同时，点出了模型的不足之处——模型可以更大，训练时间可以更长。同时提出了新的问题，给阅读者提供了未来新的研究方向——希望在视频序列中使用非常大且深的卷积网络。

第 5 章　图表分析

文章中图表的类型比较丰富，可以粗略分为视觉型插图和文字型插图，以下从宏观和微观两个角度来分析该文章的图表。

5.1　从宏观角度分析

从宏观的角度来看，我们可以从以下 4 个方面进行分析：

1. 插图的数量，文章共有 9 页，共有 9 张插图，平均每页一张插图，让读者阅读起来不会产生视觉疲劳

2. 插图的类型，文中视觉型插图（如图 5.1）有 4 张，文字型插图（公式和表格，如图 5.2）共 5 张，类型配比接近 1：1，类型均衡

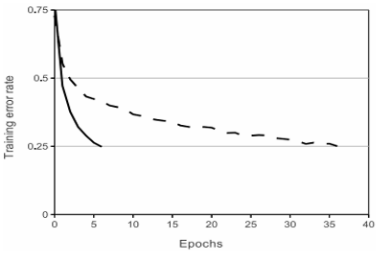


Figure 1: A four-layer convolutional neural network with ReLUs **(solid line)** reaches a 25% training error rate on CIFAR-10 six times faster than an equivalent network with tanh neurons **(dashed line)**. The learning rates for each network were chosen independently to make training as fast as possible. No regularization of any kind was employed. The magnitude of the effect demonstrated here varies with network architecture, but networks with ReLUs consistently learn several times faster than equivalents with saturating neurons.

(a)

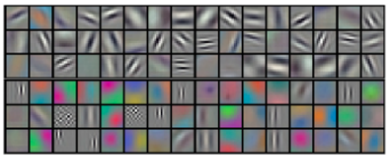


Figure 3: 96 convolutional kernels of size $11\times11\times3$ learned by the first convolutional layer on the $224\times224\times3$ input images. The top 48 kernels were learned on GPU 1 while the bottom 48 kernels were learned on GPU 2. See Section 6.1 for details.

(b)

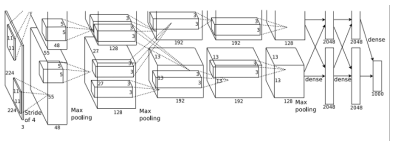


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

(c)

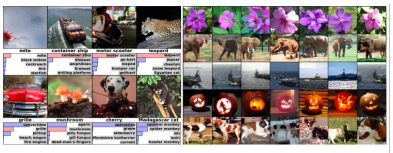
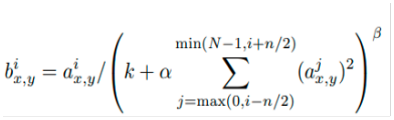


Figure 4: **(Left)** Eight ILSVRC-2010 test images and the five labels considered most probable by our model. The correct label is written under each image, and the probability assigned to the correct label is also shown with a red bar (if it happens to be in the top 5). **(Right)** Five ILSVRC-2010 test images in the first column. The remaining columns show the six training images that produce feature vectors in the last hidden layer with the smallest Euclidean distance from the feature vector for the test image.

(d)

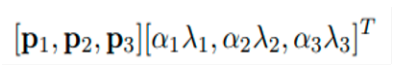
图 5.1 视觉型插图

3. 插图在文中的位置分布，文章分为 7 部分，其中第 3-6 节是主体部分，在 3.1 节、3.5 节、第 5 节、6.1 节中各有一张视觉型插图，在 3.3 节、4.1 节、第 5



$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{\beta}$$

(a)



$$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T$$

(b)

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

(c)

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

(d)

$$\begin{aligned} v_{i+1} &:= 0.9 \cdot v_i - 0.0005 \cdot \epsilon \cdot w_i - \epsilon \cdot \left\langle \frac{\partial L}{\partial w}\Big|_{w_i} \right\rangle_{D_i} \\ w_{i+1} &:= w_i + v_{i+1} \end{aligned}$$

(e)

图 5.2 文本型插图。(a) 位于 3.3 节，(b) 位于 4.1 节，(c) 位于第 6 节，(d) 位于第 6 节，（e）位于第 5 节。

节各有一个公式，第 6 节有两张表格。不难发现：图表在各节的分布也很均衡，同一小节最多不会多于 2 张插图，且除第 6 节外，每一节插图的类型都不一样，同时也没有图表的堆叠，给读者带来了良好的视觉体验，且能实时结合图表，有助于对文章内容的理解把握。

4. 内容和插图的排版，文中的插图和文字内容部分大部分很合理，除了第 6 节中的两张表格，如图 5.3所示。

其中表格的位置对文字表述部分造成了冲击，使得一个单词跨两行的现象增多，让排版变得有些混乱，一定程度上影响了读者的阅读体验。我们认为，让文字环绕在表格上下，同时将表格的标题行数减少（同时增加每行容纳的词数），带来的视觉效果更好。

6 Results

Our results on ILSVRC-2010 are summarized in Table 1. Our network achieves top-1 and top-5 test set error rates of **37.5%** and **17.0%**[5]. The best performance achieved during the ILSVRC-2010 competition was 47.1% and 28.2% with an approach that averages the predictions produced from six sparse-coding models trained on different features [2], and since then the best published results are 45.7% and 25.7% with an approach that averages the predictions of two classifiers trained on Fisher Vectors (FVs) computed from two types of densely-sampled features [24].

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

We also entered our model in the ILSVRC-2012 competition and report our results in Table 2. Since the ILSVRC-2012 test set labels are not publicly available, we cannot report test error rates for all the models that we tried. In the remainder of this paragraph, we use validation and test error rates interchangeably because in our experience they do not differ by more than 0.1% (see Table 2). The CNN described in this paper achieves a top-5 error rate of 18.2%. Averaging the predictions of five similar CNNs gives an error rate of 16.4%. Training one CNN, with an extra sixth convolutional layer over the last pooling layer, to classify the entire ImageNet Fall 2011 release (15M images, 22K categories), and then "fine-tuning" it on ILSVRC-2012 gives an error rate of 16.6%. Averaging the predictions of two CNNs that were pre-trained on the entire Fall 2011 release with the aforementioned five CNNs gives an error rate of **15.3%**. The second-best contest entry achieved an error rate of 26.2% with an approach that averages the predictions of several classifiers trained on FVs computed from different types of densely-sampled features [7].

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

Finally, we also report our error rates on the Fall 2009 version of ImageNet with 10,184 categories and 8.9 million images. On this dataset we follow the convention in the literature of using half of the images for training and half for testing. Since there is no established test set, our split necessarily differs from the splits used by previous authors, but this does not affect the results appreciably. Our top-1 and top-5 error rates on this dataset are **67.4%** and **40.9%**, attained by the net described above but with an additional, sixth convolutional layer over the last pooling layer. The best published results on this dataset are 78.1% and 60.9% [19].

图 5.3 不合理的表格排版

5.2　从微观角度分析

从微观的角度来看，可以从以下 2 个方面进行分析。

1. 视觉型插图，视觉型插图包含了图表、图示、图片，每张插图的标题都清楚的对插图进行了说明，做到了自明，同时图标、图示结构简洁清晰，图片的组合清晰美观，不得不说，该文的视觉型插图做的很棒，让读者能够更加直观的理解文章提出的 CNN 网络以及它的功能；

2. 文本型插图，其中包含了公式和表格，公式并没有大量使用，只是在论述过程中恰如其分的用来准确的表达概念，且所使用的公式形式简单，对大多数读者而言，增强了可读性，更易于理解；表格做到了自明，所表达的内容也不冗余，通过与其他模型的量化对比，让读者清晰直观的掌握到 CNN 网络的性能，但表格还有改进的空间，例如改为如表 5.1、表 5.2形式：

表 5.1 修改后的 Table 1

Model	**Top-1**	**Top-5**
Sparse coding [2]	*47.10%*	*28.20%*
SIFT+FVs [24]	*45.70%*	*25.70%*
CNN	**37.50%**	**17.00%**

Table 1: Comparison of results on ILSVRC2010 test set.

In italics are best results achieved by others

表 5.2 修改后的 Table 2

Model	**Top-1(val)**	**Top-5(val)**	**Top-5(test)**
SIFT + FVs[7]	—	—	*28.20%*
1 CNN	40.70%	18.20%	—
5 CNNs	38.10%	16.40%	**16.40%**
1 CNN*	39.00%	16.60%	—
7 CNNs*	36.70%	15.40%	**15.30%**

Table 2: Comparison of error on ILSVRC-2012 validation and test sets. In italics are best results achieved by others. Models with an asterisk* were “pre-trained” to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

第 6 章　小结

本报告对文章《ImageNet Classification with Deep Convolutional Neural Networks》进行了写作规范的研究。经过分析，我们认为该文章在写作规范的角度上是非常不错的，但是仍然有一些值得商榷、值得改进的小细节。

首先，我们认为该文章的摘要部分可以进一步改进，使得大同行也能通过文章的摘要更加理解该文章的价值。

其次，我们认为该文章的图表中，有两个表格的格式、排版不是很合理，所以我们在报告的正文部分给出了更加合理的表格格式。