

Sparse 3D convolutional neural networks

Ben Graham
b.graham@warwick.ac.uk

Department of Statistics
University of Warwick
CV4 7AL, UK

Abstract

We have implemented a convolutional neural network designed for processing sparse three-dimensional input data. The world we live in is three dimensional so there are a large number of potential applications including 3D object recognition and analysis of space-time objects. In the quest for efficiency, we experiment with CNNs on the 2D triangular-lattice and 3D tetrahedral-lattice.

1 Convolutional neural networks

Convolutional neural networks (CNNs) are powerful tools for understanding data with spatial structure such as photos. They are most commonly used in two dimensions, but they can also be applied more generally. One-dimensional CNNs are used for processing time-series such as human speech [1]. Three dimensional CNNs have been used to analyze movement in 2+1 dimensional space-time [2, 3] and for helping drones find a safe place to land [4]. Three dimensional convolutional deep belief networks have been used to recognize objects in 2.5D depth maps [5].

In [3], a *sparse* two-dimensional CNN is implemented to perform Chinese handwriting recognition. When a handwritten character is rendered at moderately high resolution on a two dimensional grid, it looks like a sparse matrix. If we only calculate the hidden units of the CNN that can actually *see* some part of the input field the pen has visited, the workload decreases. We have extended this idea to implement sparse 3D CNNs¹. Moving from two to three dimensions, the *curse of dimensionality* becomes relevant—an $N \times N \times N$ cubic grid contains many more points than an $N \times N$ square grid. However, the curse can also be taken to mean that the higher the dimension, the more likely interesting input data is to be sparse.

To motivate the idea of a sparse 3D CNN, imagine you have a loop of string with a knot in it. Mathematically, detecting and classifying knots is a hard problem; a piece of string can be very tangled without actually being knotted. Suppose you are only interested in ‘typical’ knots—humans can quite easily learn to spot the difference between, say, a trefoil knot and a figure of eight knot. If you want to take humans out of the loop, then you could train a 2D CNN to recognize and classify pictures of knots. However, pictures taken from certain angles will not contain enough information to classify the knot due to parts of the string being obscured. Suppose instead that you can trace the path of the string through three dimensional space; you could then use a 3D CNN to classify the knot. The string is essentially one dimensional, so the parts of space that the string visits will be sparse.

© 2015. The copyright of this document resides with its authors.

It may be distributed unchanged freely in print or electronic forms.

¹Software for creating sparse 2, 3 and 4 dimensional CNNs is available at <https://github.com/btgraham/SparseConvNet>

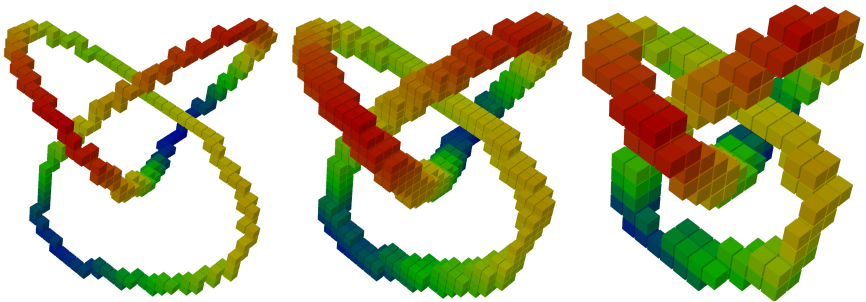


Figure 1: Left to right: A trefoil knot has been drawn in the cubic lattice; these are the input layer’s active sites. Applying a $2 \times 2 \times 2$ convolution, the number of active (i.e. non-zero) sites increases. Applying a $2 \times 2 \times 2$ pooling operation reduces the scale, which tends to decrease the number of active sites.

The example of the string is just a thought experiment. However, there are many real-world problems, in domains such as robotics and biochemistry, where understanding 3D structure is important and where sparsity is applicable.

1.1 Adding a dimension to 2D CNNs?

Recently there has been an explosion of research into conventional two-dimensional CNNs. This has gone hand-in-hand with a substantial increase in available computing power thanks to GPU computing. For photographs of size 224×224 , evaluating model C of [1]’s 19 convolutional layers requires 53 billion multiply-accumulate operations.

Although model C’s input is represented as a 3D array of size $224 \times 224 \times 3$, it is still fundamentally 2D—we can think of it as a 2D array of vectors, with each vector storing an RGB-color value. Model C’s initial convolutional layer consists of 96 convolutional filters of size 7×7 , applied with stride 2. Each filter is therefore applied $(224/2)^2$ times.

This makes 3D CNNs sound like a terrible idea. Consider adapting model C’s network architecture to accept 3D input with size $224 \times 224 \times 224 \times 3$, i.e. some kind of 3D model where each points has a color. To apply a $7 \times 7 \times 7$ convolutional filter with the same stride, we would need to apply it 112 more times than in the 2D case, with each application requiring 7 times as many operations. Extending the whole of model C to 3D would increase the computational complexity to 6.1 trillion operations. Clearly if we want to use 3D CNNs, then we need to do some things differently.

Applying the convolutions in Fourier space [2], or using separable filters [3] could help, but simply the amount of memory needed to store large 3D grids of vectors would still be a problem. Instead we try two things that work well together. Firstly we use much smaller filters, using network architectures similar to the ones introduced in [1]. The smallest non-trivial filter possible on a cubic lattice has size $2 \times 2 \times 2$, covering $2^3 = 8$ input sites. In an attempt to improve efficiency, we will also consider the tetrahedral lattice, where the smallest filter is a tetrahedron of size 2 which covers just 4 input sites. Secondly, we will only consider problems where the input is sparse. This saves us from having to have the convolutional filters visit each spatial location. If the interesting part of the input is a 1D curve or a 2D surface, then the majority of the 3D input field will receive only zero-vectors for inputs. Sparse CNNs are more efficient when used with smaller filters, as the hidden

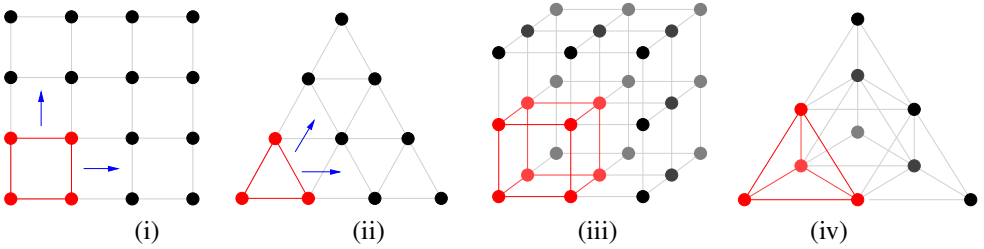


Figure 2: Convolutional filter shapes for different lattices: (i) A 4×4 square grid with a 2×2 convolutional filter. (ii) A triangular grid with size 4, and a triangular filter with size 2. (iii) A $3 \times 3 \times 3$ cubic grid, and a $2 \times 2 \times 2$ filter. (iv) A tetrahedral grid with size 3, and a filter of size 2.

layers tend to be sparser.

1.2 CNNs on different lattices

Each layer of a CNN consists of a finite graph, with a vector of input/hidden units at each site. For regular two dimensional CNNs, the graphs are square grids. The convolutional filters are square-shaped too, and they move over the underlying graph with two degrees of freedom; see Figure 2 (i). Similarly, 3D CNNs are normally defined on cubic grids. The convolutional filters are cube-shaped, and they move with three degrees of freedom; see Figure 2 (iii).

In principle we could also build 4D CNNs on hypercubic grids, and so on. However, as the dimension $d = 2, 3, 4, \dots$ increases, the size 2^d of the smallest non-trivial filter is growing exponentially. In the interests of efficiency, we will also consider CNNs with a different family of underlying graphs. In 2D, we can build CNNs based on triangles. For each layer, the underlying graph is a triangular grid, and the convolutional filters are triangular, moving with two degree of freedom; see Figure 3 (ii). In 3D, we can use a tetrahedral grid and tetrahedral filters that move with three degrees of freedom; see Figure 3 (iv). We could extend this to 4D with hypertetrahedrons, etc. In d dimensions, the smallest convolutional filters contain only $d + 1$ sites, rather than exponentially many.

To describe CNN architecture on these different lattices, we will still use the common “ $nCf/s-MPp/s\dots$ ” notation. The n counts the number of convolutional filters, f measures the linear size of the filters—the number of input sites the convolutional filter covers is f^2 , f^3 , $\binom{f+1}{2}$, $\binom{f+2}{3}$ on the square, cubic, triangular and tetrahedral lattices, respectively—and s denotes the stride. The p measures the linear size of the max-pooling regions. The $/s$ is omitted when $s = 1$ for convolutions or $s = p$ for pooling. For example, on the tetrahedral lattice $32C2 - MP3/2$ means 32 filters of size 2 which cover $\binom{2+2}{3} = 4$ input sites, followed by max pooling with pooling regions of size $\binom{3+2}{3} = 10$, and with adjacent pooling regions overlapping by one.

1.3 Sparse operations

Sparse CNNs can be thought of as an extension of the idea of sparse matrices. If a large matrix only has small number of non-zero entries per row and per column, then it makes sense to use a special data structure to store the non-zero entries and their locations; this can both dramatically reduce memory requirements and speed up operations such as matrix

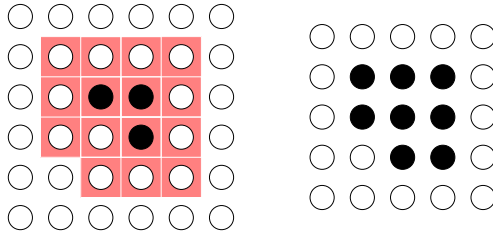


Figure 3: Calculating a 2×2 convolution for a sparse CNN: On the left is a 6×6 square grid with 3 active sites. The convolutional filter needs to be calculated at each location that covers at least on active site; this corresponds to the shaded region. The figure on the right marks the location of the eight active sites in the 5×5 output layer. Sparsity decreases with each convolution and pooling operation. However, a CNN spends most of its time processing the lower layers, so sparsity can still be useful.

multiplication. However, if 10% of the entries are non-zero, then the advantages of sparsity may be outweighed by the efficiency which dense matrix multiplication can be carried out, either using Strassen’s algorithm, or optimized GPU kernels.

The sparse CNN algorithm from [9] can be tweaked to work efficiently on general lattices. The spatial size of each of the CNN’s data layers is described by a lattice-type graph (similar to the ones in Figure 2). At each spatial location in the grid, there is a dimension-less vector of input or hidden units. Depending on the input, some of the spatial locations will be defined to be active.

- A spatial location in the input layer graph is declared *active* if the location’s vector is not the zero vector.
- Declare that a spatial location in a hidden layer is active if *any* of the spatial location in the layer below from which it receives input are active.

See Figure 3 for a 2D example of a sparse convolution, and see Figure 1 for a 3D example.

By induction, the dimension-less vectors at each non-active spatial location in the n -th hidden layer are all the same; the shared value of the vectors can be pre-computed. We will call this the *ground state* vector for the n -th level. The ground state for the input layer is just the zero vector.

We will now describe the implementation of the sparse convolution for the types of graphs shown in Figure 2. For simplicity, we will focus on case (i), the 2D square grid; the other cases are very similar.

Suppose that an image has input field size $m_{\text{in}} \times m_{\text{in}}$, and that the number of active spatial locations is $a_{\text{in}} \in \{0, 1, \dots, m_{\text{in}}^2\}$. Suppose an $f \times f$ convolutional filter will act on the image, and let n_{in} and n_{out} denote the number of input and output features per spatial location. The input to the first convolutional operation consists of:

- A matrix M_{in} with size $a_{\text{in}} \times n_{\text{in}}$. Each row corresponds to the vector at one of the active spatial locations.
- A map or hash table H_{in} of *(key,value)* pairs. The *keys* are the active spatial locations. The *values* record the number of the corresponding row in M_{in} .
- The input layer’s ground state vector g_{in} .

- An $(f^2 n_{\text{in}}) \times n_{\text{out}}$ matrix W containing the weights that define the convolution.
- A vector B of length n_{out} specifying the values of the bias units.

To calculate the output of the first hidden layer:

1. Iterate through H_{in} and determine the number a_{out} of active spatial locations in the output layer. A site in the output layer is active if any of the input sites are active. Build a hash table H_{out} to uniquely identify each of the active output spatial locations with one of the integers $1, 2, \dots, a_{\text{out}}$.
2. Use H_{in} , M_{in} and g_{in} to build a matrix Q of size $a_{\text{out}} \times (f^2 n_{\text{in}})$; each row of Q should correspond to the inputs visible to the convolutional filter at the corresponding output spatial location.
3. Calculate $M_{\text{out}} = Q \times W + B$.

We implemented step 1 on the CPU and steps 2 and 3 on the GPU. If W is small, the computational bottleneck will be I/O-related, steps 1 and 2. If W is large, the bottleneck will be performing the dense matrix multiplication in step 3 to calculate M_{out} .

The procedure for max-pooling is similar. Max-pooling is always I/O-bound.

2 Experiments

We have performed experiments to test triangular and sparse 3D CNNs. Unlike the 2D case, there are not yet any standard benchmarks for evaluating 3D CNNs, so we just picked a range of different types of data. When faced with a trade-off between computational cost and accuracy, we have preferred to train smaller network to see what can be achieved on a limited computational budget, rather than trying to maximize performance at any cost.

For some of the experiments we used n -fold repetitive testing: we processed each test case n times, with some form of data augmentation, for n a small integer, and averaged the output.

2.1 Square versus triangular 2D convolutions

As a sanity test regarding our unusually shaped CNNs, we first did a 2D experiment with the CIFAR-10 dataset of small pictures [8] to compare CNNs on the square and triangular lattices. We will call the networks SquareNet and TriangLeNet, respectively. Both networks have 12 small convolutional layers split into pairs by 5 layers of max-pooling, and with the n -th pair of convolutional filters each having $32n$ output features:

$$32C2 - 32C2 - MP3/2 - \dots - MP3/2 - 192C2 - 192C2 - \text{output}$$

We extended the training data using affine transformations. For the triangular lattice, we converted the images to triangular coordinates using an additional affine transformation. See Table 1 for the results.

TriangLeNet has a computational cost that is 26% lower than the more conventional SquareNet. In terms of test errors, there does not seem to be any real difference between the two networks.

CNN	MegaOps	test error	12-fold test error
SquareNet	41	9.24%	7.66%
TriangLeNet	30	9.70%	7.50%

Table 1: Comparison between square and triangular 2D CNNs for CIFAR-10

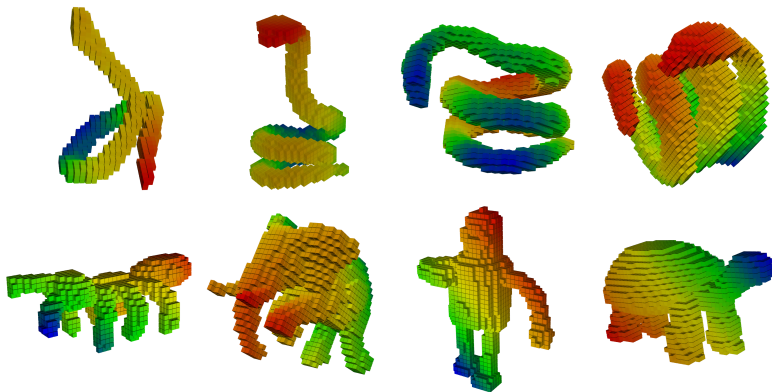


Figure 4: Items from the 3D object dataset used in Section 2.2, embedded into a $40 \times 40 \times 40$ cubic grid. Top row: four items from the snake class. Bottom row: an ant, an elephant, a robot and a tortoise.

2.2 Object recognition

To test the 3D CNN, we used a dataset of 3D objects², each stored as a mesh of triangles in the OFF-file format. The dataset contains 1200 exemplars split evenly between 50 classes (aliens, ants, armadillo, ...). The dataset was intended to be used for unsupervised learning, but as CNNs are most often used for supervised learning, we used 6-fold cross-validation to measure the ability of our 3D CNNs to learn shapes. To stop the dataset being too easy, we randomly rotated the objects during training and testing. This is to force the CNN to truly learn to recognize shape, and not rely on some classes of objects tending to have a certain orientation.

All the CNNs we tested took the form

$$32C2 - \text{pooling} - 64C2 - \text{pooling} - 96C2 - \dots - \text{output}.$$

We rendered the 3D models at a variety of different scales, and varied the number of levels of pooling accordingly. We tried using MP3/2 pooling on the cubic and tetrahedral lattices. We also tried a stochastic form of max-pooling on the cubic lattice which we denote FMP[\square]; we used FMP to downsample the hidden layer by a factor of $2^{2/3} \approx 1.59$; this allows us to gently increase the number of learnt layers for a given input scale. See Figure 5.

The tetrahedral CNNs are substantially cheaper computationally, but less accurate at the smallest scale. The FMP pooling provides the highest accuracy when the scale is small, but they are quite a bit more expensive. If we look at the number of test samples that can be processed per second, we see that for such small CNNs the calculations are actually I/O-bound, so tetrahedral network is not as much faster as we might have expected based on the

²SHREC2015 Non-rigid 3D Shape Retrieval dataset <http://www.icst.pku.edu.cn/zliang/shrec15-non-rigid/data.html>

scale	pooling	$\times 10^6$ operations	tests/s
20	4 \times MP3/2 \triangle	9	1133
40	5 \times MP3/2 \triangle	35	675
80	6 \times MP3/2 \triangle	143	286
20	4 \times MP3/2 \square	36	1190
40	5 \times MP3/2 \square	126	794
80	6 \times MP3/2 \square	406	310
20	6 \times FMP \square	116	1100
32	7 \times FMP \square	279	849

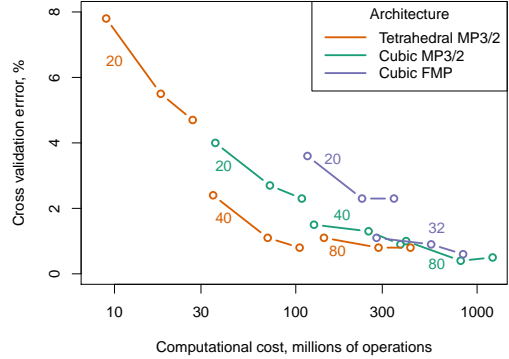


Figure 5: 6-fold cross-validation error rate for 3D object recognition for different CNN architectures. The lines in the graph correspond to performing 1-, 2- and 3-fold testing with a given CNN. The table given the computational complexity and speed of the network on a Nvidia GeForce GTX 780 GPU.

computational cost. However, with a less powerful processor, it is likely that there would be a speed advantage to the tetrahedral lattice.

2.3 2D space + 1D time = 3D space-time

The CASIA-OLHWDB1.1 database contains online handwriting samples of the 3755 GBK level-1 Chinese characters [14]. There are approximately 240 training characters, and 60 test characters, per class. *Online* means that the pen strokes were recorded in the order they were made.

A test error of 5.61% is achieved by drawing the characters with size 40×40 and learning to recognize their pictures with a 2D CNN [14]. Evaluating that network’s four convolutional layers requires 72 million multiply-accumulate operations.

With a 3D CNN, we can use the order in which the strokes were written to represent each character as a collection of paths in 2+1 dimensional space-time with size $40 \times 40 \times 40$. A 3D CNN with architecture

$$32C3 - MP3/2 - 64C2 - MP3/2 - 128C2 - MP3/2 - 256C2 - MP3/2 - 512C3 - \text{output}$$

requires on average 118 million operations to evaluate, and produced a test error of 4.93%. We deliberately kept the input spatial size the same, so any improvements would be due to the introduction of the time dimension.

2.4 Human action recognition

Recognizing actions in videos is another example of a 2+1 dimensional space-time problem. A simple way of turning a video into a sparse 3D object is to take the difference between successive frames, and then setting to zero any values with absolute value below some threshold. We tried this approach on two datasets, the simpler RHA dataset³ [14] with 6 classes of actions, and the harder UCF101 [14] dataset. We scaled the UCF101 video down by 50% to

³<http://www.nada.kth.se/cvap/actions/>



Figure 6: An image from the two video datasets used in Section 2.4, and the difference between that frame and the previous frame.

have the same size as the HRA videos, 160×120 . In both cases we used a cubic CNN:

$$32C2 - MP3/2 - 64C2 - MP3/2 - \dots - 192C2 - MP3/2 - 224C2 - \text{output}.$$

For RHA, (mean) accuracy of 71.7% is reported in [14]. We used a threshold of 12%, which resulted in 1.7% of input pixels being active. Our approach yielded 88.0% accuracy with a computational cost of 1.1 billion operations per test case.

For UCF101, accuracy of 43.90% is reported in [1]. We used a threshold of 13%, which resulted in 3.1% of input pixels being active. The computational cost was higher than for RHA, 2.7 billion operations, as the videos are more complicated. Single testing produced an accuracy of 60.4%, rising to 67.8% with 12-fold testing.

These results are not state of the art. However, they do seem to strike a good balance in terms of computational cost. Also, we have not done any work to try to optimize our results. There are different ways of encoding a video’s ‘optical flow’ that we have not had a chance to explore yet.

3 Conclusion

We have shown that sparse 3D CNNs can be implemented efficiently, and produce interesting results for a variety of types of 3D data. There are potential applications that we have not yet tried. In biochemistry, there are large databases of 3D molecular structure. Proteins that are encoded differently may fold to produce similar shapes with similar functions. In robotics, it is natural to build 3D models by combining one or more 2D images with depth detector databases. Sparse 3D CNNs could be used to analyse these models.

References

- [1] D. Ciresan, U. Meier, and J. Schmidhuber. [Multi-column deep neural networks for image classification](#). In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649, 2012. doi: 10.1109/CVPR.2012.6248110.
- [2] Ben Graham. Fractional max-pooling, 2014. <http://arxiv.org/abs/1412.6071>.
- [3] Ben Graham. Spatially-sparse convolutional neural networks. 2014. URL <http://arxiv.org/abs/1409.6070>.

- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2014. <http://arxiv.org/abs/1502.01852>.
- [5] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, January 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.59. URL <http://dx.doi.org/10.1109/TPAMI.2012.59>.
- [6] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.
- [7] Amir Roshan Zamir Khurram Soomro and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. Technical report, November 2012. CRCV-TR-12-01.
- [8] Alex Krizhevsky. [Learning Multiple Layers of Features from Tiny Images](#). Technical report, 2009.
- [9] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.
- [10] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang. CASIA online and offline Chinese handwriting databases. In *Proc. 11th International Conference on Document Analysis and Recognition (ICDAR), Beijing, China*, pages 37–41, 2011. URL <http://www.nlpr.ia.ac.cn/databases/download/ICDAR2011-CASIA%20databases.pdf>.
- [11] Michael Mathieu, Mikael Henaff, and Yann LeCun. Fast training of convolutional networks through ffts. In *International Conference on Learning Representations (ICLR2014)*. CBLS, April 2014. URL <http://openreview.net/document/aa6ab717-ca19-47e1-a958-823b9a106ca9>.
- [12] D. Maturana and S. Scherer. 3D Convolutional Neural Networks for Landing Zone Detection from LiDAR. In *ICRA*, 2015.
- [13] Roberto Rigamonti, Amos Sironi, Vincent Lepetit, and Pascal Fua. Learning separable filters. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013.
- [14] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A Local SVM approach, 2004. URL <http://dx.doi.org/10.1109/ICPR.2004.747>.
- [15] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.