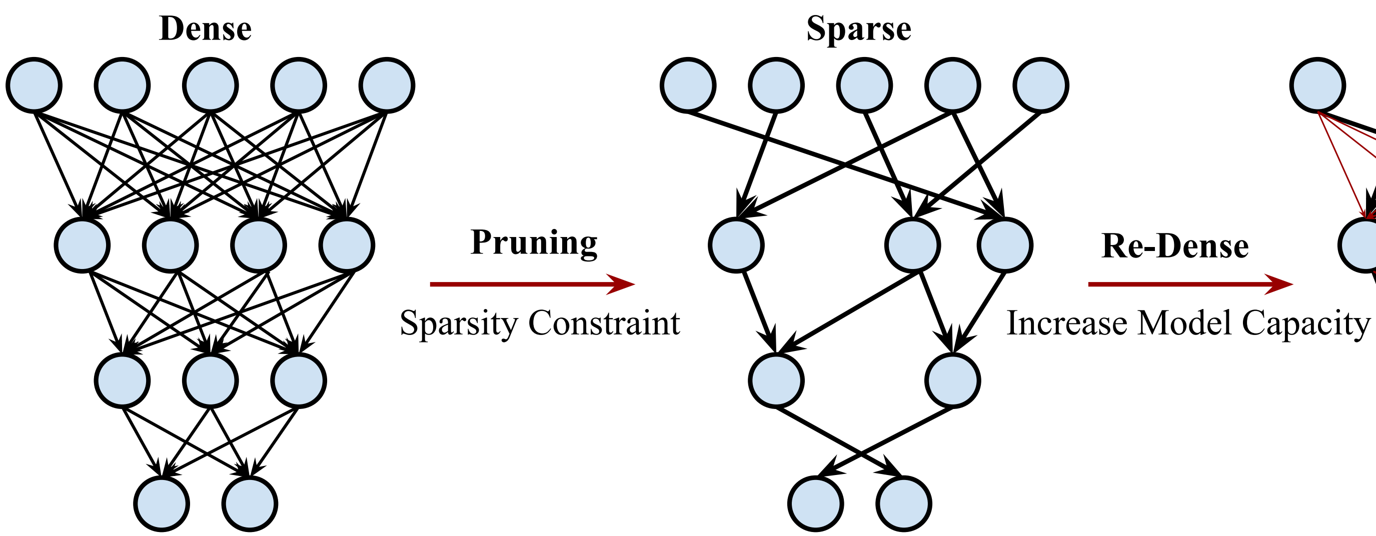
DSD: Dense-Sparse-Dense Training for Deep Neural Networks

Song Han[1], Jeff Pool[2], Sharan Narang[3]
Huizi Mao[1], Enhao Gong[1], Shijian Tang[1], Erich Elsen[3], Peter Vajda[4], Manohar Paluri[4]
John Tran[2], Bryan Catanzaro[2], William J. Dally[1,2]
[1] Stanford University [2] NVIDIA [3] Baidu [4] Facebook

Dense-Sparse-Dense Training



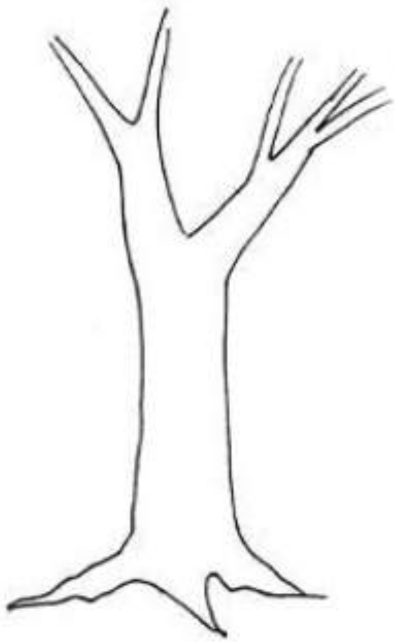
DSD produces same model architecture but can find better optimization solution, arrives at better local minima, and achieves higher prediction accuracy across a wide range of deep neural networks on CNNs / RNNs / LSTMs.

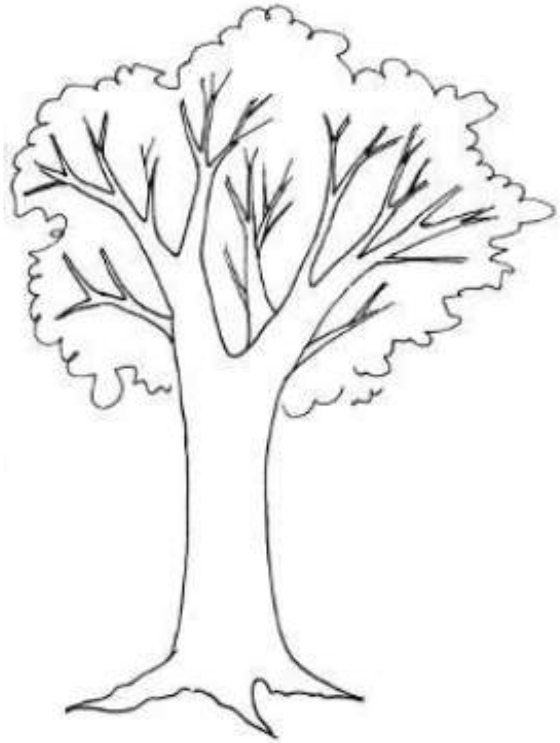
Han et al "DSD: Dense-Sparse-Dense Training for Deep Neural Networks", ICLR'17

DSD: Motivation

- Deep model compression without losing accuracy means there's significant redundancy
- Which shows the inadequacy of current training methods.
- Dense-Sparse-Dense training provides a strong regularization to make current models converge at higher accuracy.

Intuition





Learn the trunk first

Then learn the leaves

Dense-Sparse-Dense Training Flow

1. **Train a network.**

   (Or download one from the model zoo.)

2. **Prune the network.**

   (Set the smallest 30%-50% parameters to be zero.)

3. **Fine-tune the sparse network.**

   (Train with the sparsity mask, recover the accuracy.)

4. **Remove the sparsity constraint.**

   (Throw away the sparsity mask.)

5. **Fine-tune the dense network.**

   (Watch the accuracy increase even more.)

Algorithm

Algorithm 1: Workflow of DSD training

Initialization: $W^{(0)}$ *with* $W^{(0)} \sim N(0, \Sigma)$
Output : $W^{(t)}$.

——————————————————————— *Initial Dense Phase* ———

while *not converged* **do**
$\tilde{W}^{(t)} = W^{(t-1)} - \eta^{(t)} \nabla f(W^{(t-1)}; x^{(t-1)})$;
$t = t + 1$;
end

——————————————————————— *Sparse Phase* ———

// initialize the mask by sorting and keeping the Top-k weights.
$S = sort(|W^{(t-1)}|); \quad \lambda = S_{k_i}; \quad Mask = \mathbb{1}(|W^{(t-1)}| > \lambda)$;
while *not converged* **do**
$\tilde{W}^{(t)} = W^{(t-1)} - \eta^{(t)} \nabla f(W^{(t-1)}; x^{(t-1)})$;
$\tilde{W}^{(t)} = W^{(t)} \cdot Mask$;
$t = t + 1$;
end

——————————————————————— *Final Dense Phase* ———

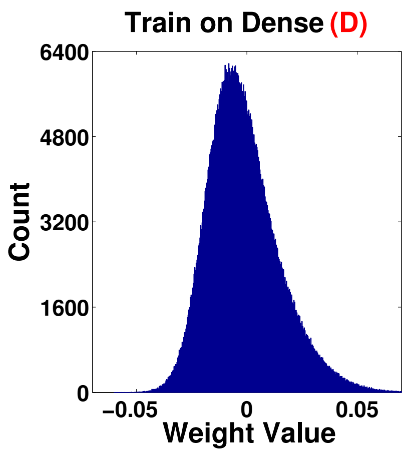
while *not converged* **do**
$\tilde{W}^{(t)} = W^{(t-1)} - \eta^{(t)} \nabla f(W^{(t-1)}; x^{(t-1)})$;
$t = t + 1$;
end
goto *Sparse Phase* for iterative DSD;

Related Work

- **Dropout** [1] **and DropConnect** [2]
  - Dropout use a *random* sparsity pattern.
  - DSD training learns with a *deterministic* data driven sparsity pattern.
- **Distillation** [3]
  - Transfer the knowledge from the large model to a small model.
  - Both DSD and Distillation don’t incur architectural changes.
- **Simulated Annealing** [4]
  - *Randomly* jumps with decreasing probability on the search graph.
  - DSD *deterministically* deviates from the converged solution by removing the *small* weights and enforcing a sparsity support.
  - Both could be done iteratively.

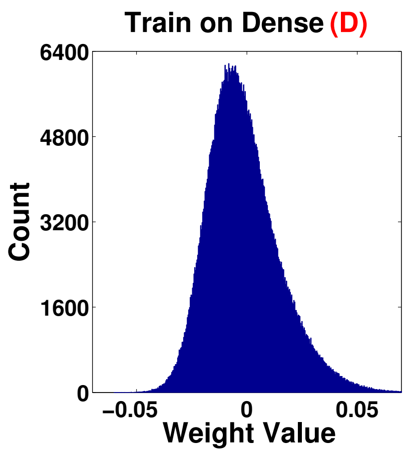
[1] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research* 15.1 (2014): 1929-1958.
[2] Wan, Li, et al. "Regularization of neural networks using dropconnect." *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013.
[3] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).
[4] Chii-Ruey Hwang. Simulated annealing: theory and applications. Acta Applicandae Mathematicae, 12(1): 108–111, 1988.

Weight Distribution

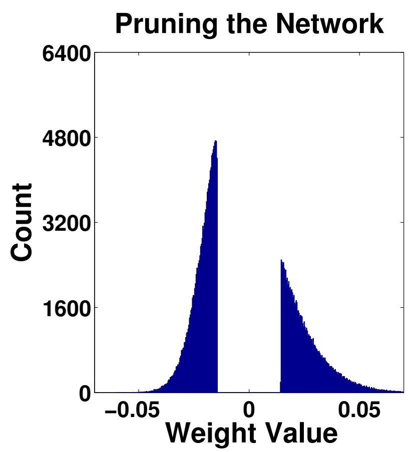


(a)

Weight Distribution

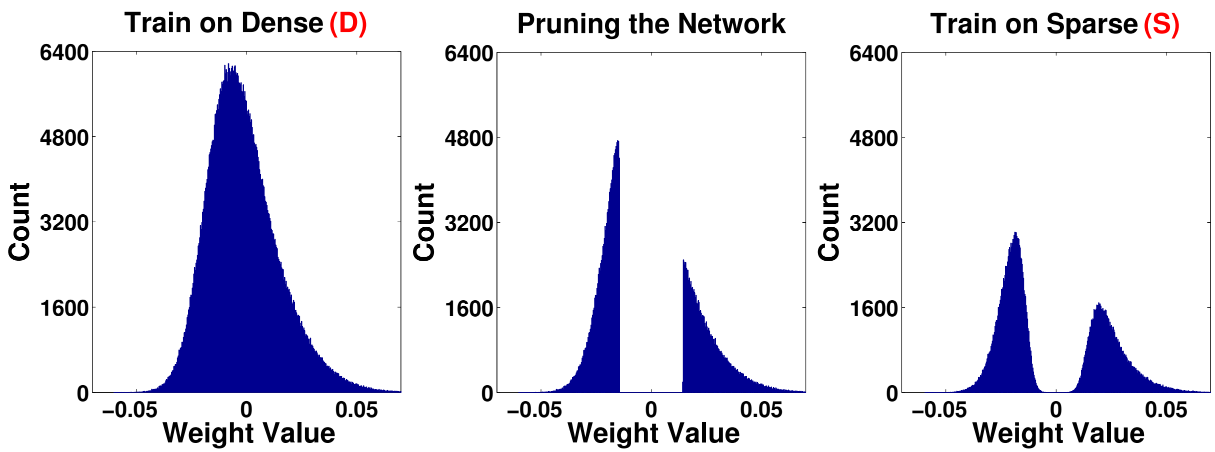


(a)



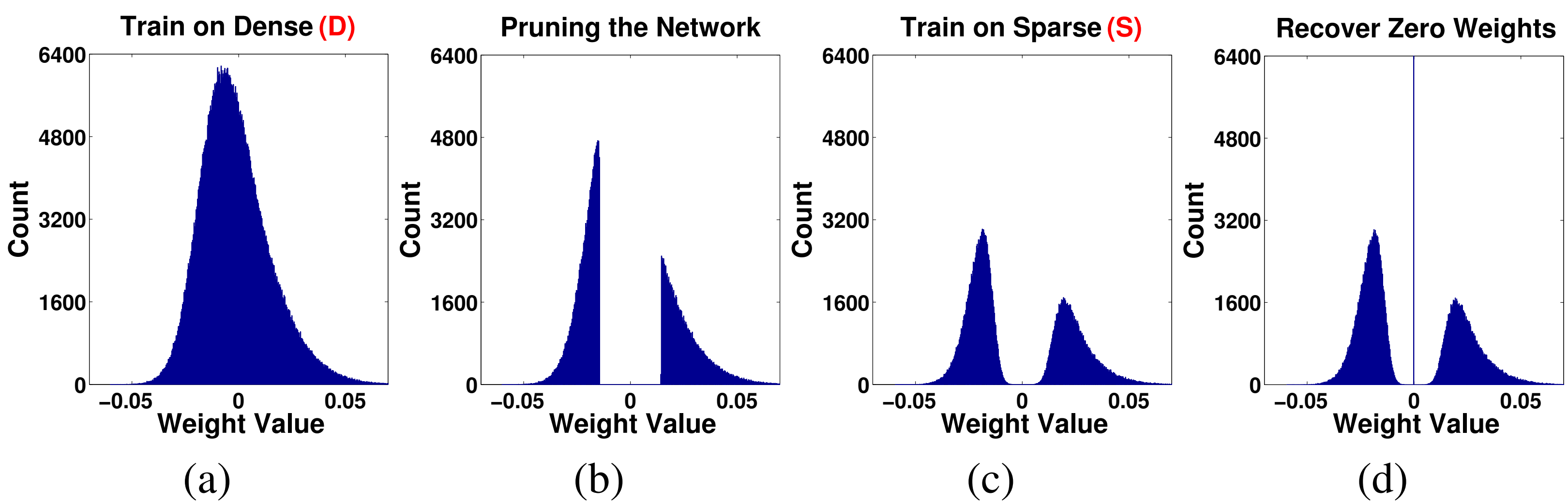
(b)

Weight Distribution

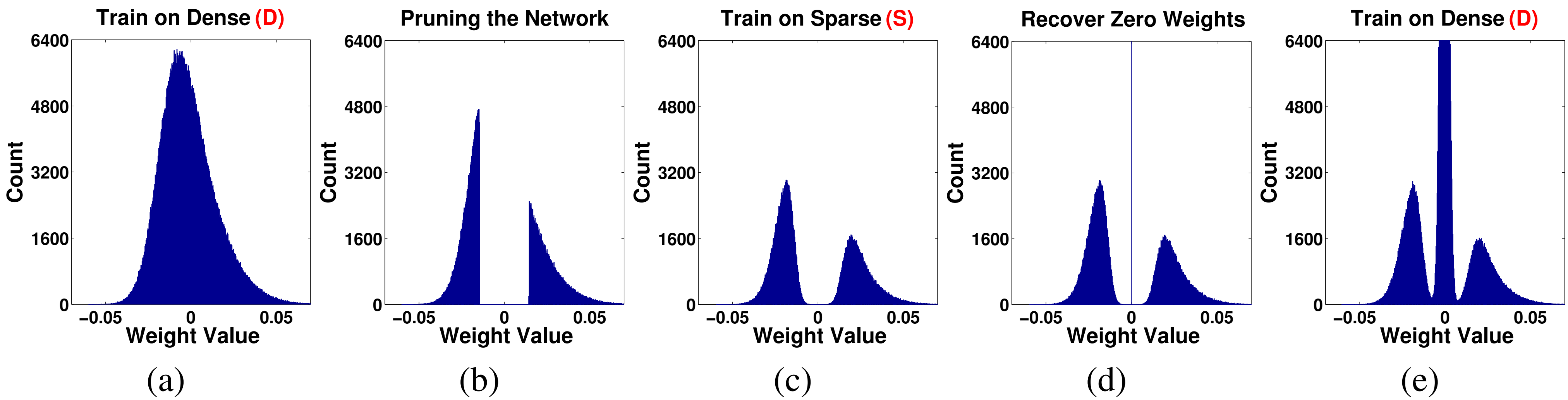


(a) (b) (c)

Weight Distribution



Weight Distribution



DSD is General Purpose: CNN, RNN, LSTM

Network	Domain	Dataset	Type	Baseline		DSD	Abs. Imp.	Rel. Imp.
GoogleNet	Vision	ImageNet	CNN	31.1%	→	**30.0%**	1.1%	3.6%
VGG-16	Vision	ImageNet	CNN	31.5%	→	**27.2%**	4.3%	13.7%
ResNet-18	Vision	ImageNet	CNN	30.4%	→	**29.3%**	1.1%	3.7%
ResNet-50	Vision	ImageNet	CNN	24.0%	→	**23.2%**	0.9%	3.5%

Open Sourced DSD Model Zoo: https://songhan.github.io/DSD

The beseline results of AlexNet, VGG16, GoogleNet, SqueezeNet are from Caffe Model Zoo. ResNet18, ResNet50 are from fb.resnet.torch.

[Han et al. ICLR 2017]

DSD is General Purpose: CNN, RNN, LSTM

Network	Domain	Dataset	Type	Baseline		DSD	Abs. Imp.	Rel. Imp.
GoogleNet	Vision	ImageNet	CNN	31.1%	→	**30.0%**	1.1%	3.6%
VGG-16	Vision	ImageNet	CNN	31.5%	→	**27.2%**	4.3%	13.7%
ResNet-18	Vision	ImageNet	CNN	30.4%	→	**29.3%**	1.1%	3.7%
ResNet-50	Vision	ImageNet	CNN	24.0%	→	**23.2%**	0.9%	3.5%
NeuralTalk	Caption	Flickr-8K	LSTM	16.8	→	**18.5**	1.7	10.1%

Open Sourced DSD Model Zoo: https://songhan.github.io/DSD

The beseline results of AlexNet, VGG16, GoogleNet, SqueezeNet are from Caffe Model Zoo. ResNet18, ResNet50 are from fb.resnet.torch.

[Han et al. ICLR 2017]

DSD is General Purpose: CNN, RNN, LSTM

Network	Domain	Dataset	Type	Baseline		DSD	Abs. Imp.	Rel. Imp.
GoogleNet	Vision	ImageNet	CNN	31.1%	→	**30.0%**	1.1%	3.6%
VGG-16	Vision	ImageNet	CNN	31.5%	→	**27.2%**	4.3%	13.7%
ResNet-18	Vision	ImageNet	CNN	30.4%	→	**29.3%**	1.1%	3.7%
ResNet-50	Vision	ImageNet	CNN	24.0%	→	**23.2%**	0.9%	3.5%
NeuralTalk	Caption	Flickr-8K	LSTM	16.8	→	**18.5**	1.7	10.1%
DeepSpeech	Speech	WSJ'93	RNN	33.6%	→	**31.6%**	2.0%	5.8%
DeepSpeech-2	Speech	WSJ'93	RNN	14.5%	→	**13.4%**	1.1%	7.4%

Open Sourced DSD Model Zoo: https://songhan.github.io/DSD

The beseline results of AlexNet, VGG16, GoogleNet, SqueezeNet are from Caffe Model Zoo. ResNet18, ResNet50 are from fb.resnet.torch.

Multiple DSD Iterations

DeepSpeech2 Network	Baseline WER		DSD Iter 1 WER		DSD Iter 2 WER	Abs. Accuracy Improve	Rel. Error Reduction
WSJ '92	9.55	⟶	9.11	⟶	9.02	0.53	**5.6%**
WSJ '93	14.52	⟶	13.96	⟶	13.44	1.08	**7.4%**

Results on Caption Generation

✗ **Baseline**: a boy in a red shirt is climbing a rock wall.

✗ **Sparse**: a young girl is jumping off a tree.

✓ **DSD**: a young girl in a pink shirt is swinging on a swing.

○ **Baseline**: a basketball player in a red uniform is playing with a ball.

○ **Sparse**: a basketball player in a blue uniform is jumping over the goal.

✓ **DSD**: a basketball player in a white uniform is trying to make a shot.

✓ **Baseline**: two dogs are playing together in a field.

✓ **Sparse**: two dogs are playing in a field.

✓ **DSD**: two dogs are playing in the grass.

✗ **Baseline**: a man and a woman are sitting on a bench.

○ **Sparse**: a man is sitting on a bench with his hands in the air.

○ **DSD**: a man is sitting on a bench with his arms folded.

✗ **Baseline**: a person in a red jacket is riding a bike through the woods.

✓ **Sparse**: a car drives through a mud puddle.

✓ **DSD**: a car drives through a forest.

Table 7: DSD results on NeuralTalk

NeuralTalk	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Sparsity
Baseline	57.2	38.6	25.4	16.8	0%
Sparse	58.4	39.7	26.3	17.5	80%
DSD	**59.2**	**40.7**	**27.4**	**18.5**	0%
Improvement (abs)	2.0	2.1	2.0	1.7	-
Improvement (rel)	**3.5%**	**5.4%**	**7.9%**	**10.1%**	-

BLEU score baseline from Neural Talk model zoo by Andrej Karpathy

Results on Caption Generation

Baseline: a boy is swimming in a pool.
Sparse: a small black dog is jumping into a pool.
DSD: a black and white dog is swimming in a pool.

Baseline: a group of people are standing in front of a building.
Sparse: a group of people are standing in front of a building.
DSD: a group of people are walking in a park.

Baseline: two girls in bathing suits are playing in the water.
Sparse: two children are playing in the sand.
DSD: two children are playing in the sand.

Baseline: a man in a red shirt and jeans is riding a bicycle down a street.
Sparse: a man in a red shirt and a woman in a wheelchair.
DSD: a man and a woman are riding on a street.

Baseline: a group of people sit on a bench in front of a building.
Sparse: a group of people are standing in front of a building.
DSD: a group of people are standing in a fountain.

Baseline: a man in a black jacket and a black jacket is smiling.
Sparse: a man and a woman are standing in front of a mountain.
DSD: a man in a black jacket is standing next to a man in a black shirt.

Baseline: a group of football players in red uniforms.
Sparse: a group of football players in a field.
DSD: a group of football players in red and white uniforms.

Baseline: a dog runs through the grass.
Sparse: a dog runs through the grass.
DSD: a white and brown dog is running through the grass.

Results on Caption Generation

Baseline: a man in a red shirt is standing on a rock.
Sparse: a man in a red jacket is standing on a mountaintop.
DSD: a man is standing on a rock overlooking the mountains.

Baseline: a group of people are sitting in a subway station.
Sparse: a man and a woman are sitting on a couch.
DSD: a group of people are sitting at a table in a room.

Baseline: a man in a red jacket is standing in front of a white building.
Sparse: a man in a black jacket is standing in front of a brick wall.
DSD: a man in a black jacket is standing in front of a white building.

Baseline:a young girl in a red dress is holding a camera.
Sparse: a little girl in a pink dress is standing in front of a tree.
DSD: a little girl in a red dress is holding a red and white flowers.

Baseline: a soccer player in a red and white uniform is playing with a soccer ball.
Sparse: two boys playing soccer.
DSD: two boys playing soccer.

Baseline: a girl in a white dress is standing on a sidewalk.
Sparse: a girl in a pink shirt is standing in front of a white building.
DSD: a girl in a pink dress is walking on a sidewalk.

Baseline: a young girl in a swimming pool.
Sparse: a young boy in a swimming pool.
DSD: a girl in a pink bathing suit jumps into a pool.

Baseline: a soccer player in a red and white uniform is running on the field.
Sparse: a soccer player in a red uniform is tackling another player in a white uniform.
DSD: a soccer player in a red uniform kicks a soccer ball.

Results on Caption Generation

Baseline: a man in a red shirt is sitting in a subway station.
Sparse: a woman in a blue shirt is standing in front of a store.
DSD: a man in a black shirt is standing in front of a restaurant.

Baseline: a surfer is riding a wave.
Sparse: a man in a black wetsuit is surfing on a wave.
DSD: a man in a black wetsuit is surfing a wave.

Baseline: two young girls are posing for a picture.
Sparse: a young girl with a blue shirt is blowing bubbles.
DSD: a young boy and a woman smile for the camera.

Baseline: a snowboarder flies through the air.
Sparse: a person is snowboarding down a snowy hill.
DSD: a person on a snowboard is jumping over a snowy hill.

Baseline: a man in a red shirt is standing on top of a rock.
Sparse: a man in a red shirt is standing on a cliff overlooking the mountains.
DSD: a man is standing on a rock overlooking the mountains.

Baseline: a group of people sit on a bench.
Sparse: a group of people are sitting on a bench.
DSD: a group of children are sitting on a bench.

Baseline: a little boy is playing with a toy.
Sparse: a little boy in a blue shirt is playing with bubbles.
DSD: a baby in a blue shirt is playing with a toy.

Baseline: a brown dog is running through the grassy.
Sparse: a brown dog is playing with a ball.
DSD: a brown dog is playing with a ball.

Results on Caption Generation

Baseline: a boy in a red shirt is jumping on a trampoline.
Sparse: a boy in a red shirt is jumping in the air.
DSD: a boy in a red shirt is jumping off a swing.

Baseline: a man is standing on the edge of a cliff.
Sparse: a man is standing on the shore of a lake.
DSD: a man is standing on the shore of the ocean.

Baseline: two people are riding a boat on the beach.
Sparse: two people are riding a wave on a beach.
DSD: a man in a yellow kayak is riding a wave.

Baseline: a black and white dog is running on the beach.
Sparse: a black and white dog running on the beach.
DSD: a black dog is running on the beach.

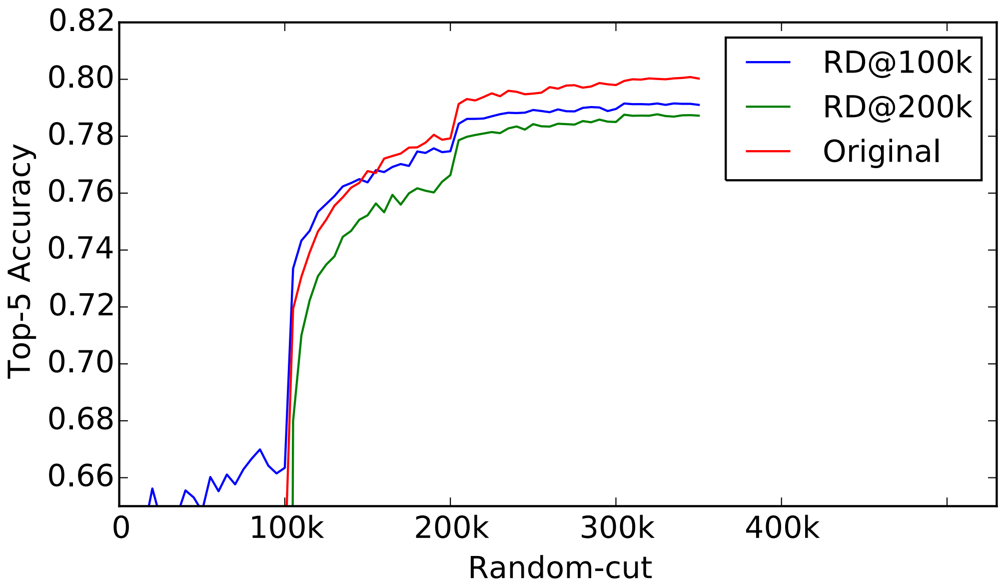
Baseline: a man and a dog are playing with a ball.
Sparse: a man and a woman are playing tug of war.
DSD: a man and a woman are playing with a dog.

Baseline: a group of people are standing in a room.
Sparse: a group of people gather together.
DSD: a group of people are posing for a picture.

Baseline: a man in a red jacket is riding a bike through the woods.
Sparse: a man in a red jacket is doing a jump on a snowboard.
DSD: a person on a dirt bike jumps over a hill.

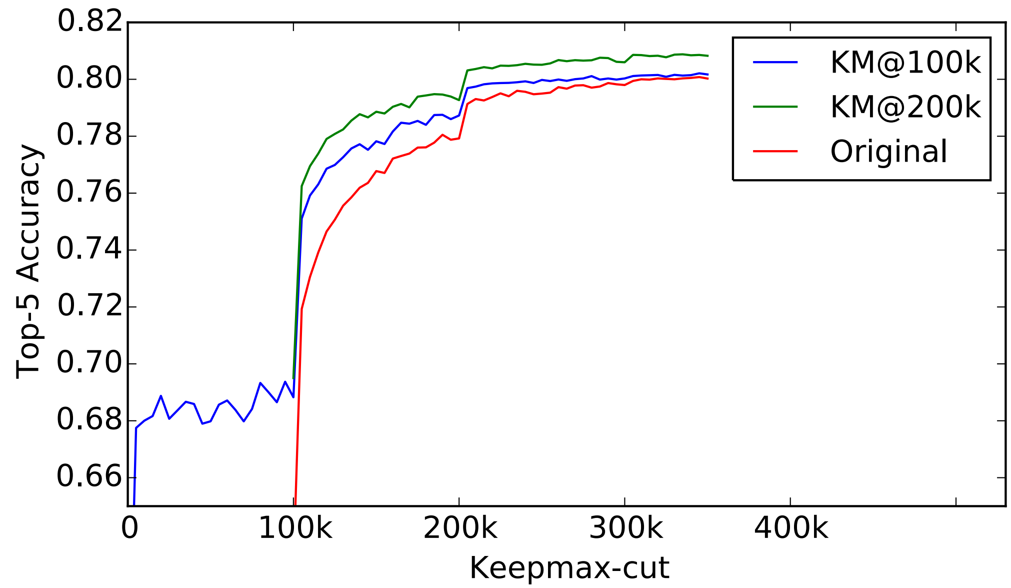
Baseline: a man in a red jacket and a helmet is standing in the snow.
Sparse: a man in a red jacket and a helmet is standing in the snow.
DSD: a man in a red jacket is standing in front of a snowy mountain.

Hypothesis: Sparsity Pattern Encode Knowledge



Random sparsity

Worse accuracy



Truncation-based sparsity

Better accuracy

Han et al. “DSD: Regularizing Deep Neural Networks with Dense-Sparse-Dense Training ”, arXiv 2016

Why DSD Works

- DSD perturbs the learning dynamics and allows the network to jump away from saddle points.
- The sparsity constraint moves the optimization to a lower-dimensional space where the loss surface is smoother and tend to be more robust to noise.
- DSD gives the optimization a second (or more) chance during the training process to re-initialize
- DSD breaks symmetry

DSD in Practice

- There’s only **a single hyper parameter** to tune, which is the sparsity.
- Empirical sparsity: 30%-50%
- Uniform sparsity for each layer except the first layer.
- No aggressive pruning needed:
- No iterative pruning needed.
- No need to determine threshold.
- The epochs are decided when it converges.

DSD Model Zoo

DSD model zoo. Better accuracy models from DSD training on Imagenet with same model architecture.

</> View DSD Model Zoo on GitHub

zip Download

tar Download

DSD Model Zoo

This repo contains pre-trained models by Dense-Sparse-Dense(DSD) training on Imagenet.

Compared to conventional training method, dense→sparse→dense (DSD) training yielded higher accuracy with same model architecture.

Sparsity is a powerful form of regularization. Our intuition is that, once the network arrives at a local minimum given the sparsity constraint, relaxing the constraint gives the network more freedom to escape the saddle point and arrive at a higher-accuracy local minimum.

Open Sourced DSD Model Zoo: https://songhan.github.io/DSD

Conclusion

- We introduce DSD, a dense-sparse-dense training framework that regularizes neural networks by pruning and then restoring connections.

- DSD training achieves superior optimization performance. Our numerical results and empirical tests show the inadequacy of current training methods for which we have provided an effective solution.