ICML 2016
THE 33rd INTERNATIONAL CONFERENCE ON MACHINE LEARNING

University of Michigan

Augmenting Supervised Neural Networks with Unsupervised Objectives for Large-scale Image Classification

Yuting Zhang, Kibok Lee, Honglak Lee

University of Michigan, Ann Arbor

Unsupervised and supervised deep learning

- Deep feature representations can be learned in supervised and unsupervised manners.
  - Supervised objectives learns from the correspondence between data and label space.
  - Unsupervised objectives learns from the data space itself.

- Supervised deep learning
  - Deep neural networks, convolutional neural networks, recurrent neural networks, …
  - Task-specific, requires large amounts of supervision

- Unsupervised deep learning
  - Stacked autoencoders, deep belief networks, deep Boltzmann machines, …
  - Preserves input information, can leverage large amounts of unlabeled data, but may be suboptimal for supervised tasks.

Yuting Zhang, Kibok Lee, Honglak Lee

Unsupervised and supervised deep learning

- Historically, unsupervised learning (e.g., SAE) can be used as a pretraining step for improving and even enabling the supervised learning of deep networks.

- However, such pretraining became unnecessary if the deep neural network is initialized properly, and large amount of labeled data are available.
  - E.g., large-scale convolutional neural networks: AlexNet (Krizhevsky et al., 2012), VGGNet (Simounyan and Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), etc.

- As a result, unsupervised deep learning has been overshadowed by supervised methods.

Revisiting the importance of unsupervised learning

- Pretraining: Unsupervised → Supervised

Yuting Zhang, Kibok Lee, Honglak Lee

Revisiting the importance of unsupervised learning

- **Combination:** Unsupervised + Supervised

  reconstruction + classification

- Previous work:
  - Autoencoders: Ranzato & Szummer (2008); Larochelle et al. (2009)
  - (Restricted) Boltzmann machines: Larochelle & Bengio, (2008); Goodfellow et al. (2013); Sohn et al. (2013)
  - Dictionary learning: Boureau et al. (2010); Mairal et al. (2010)

Revisiting the importance of unsupervised learning

- **Combination:** Unsupervised + Supervised
  reconstruction + classification

- Previous work:
  - Autoencoders: Ranzato & Szummer (2008); Larochelle et al. (2009)
  - (Restricted) Boltzmann machines: Larochelle & Bengio, (2008); Goodfellow et al. (2013); Sohn et al. (2013)
  - Dictionary learning: Boureau et al. (2010); Mairal et al. (2010)
  - **Ladder network**: Rasmus et al. (2015)
    - layer-wise skip links & pathway combinators
  - Stacked "what-where" AE (**SWWAE**): Zhao et al. (2015)
    - using unpooling switches (Zeiler and Fergus, 2009)

- Promising for improving classification performance, but have not been shown to be beneficial for large-scale supervised deep neural nets.

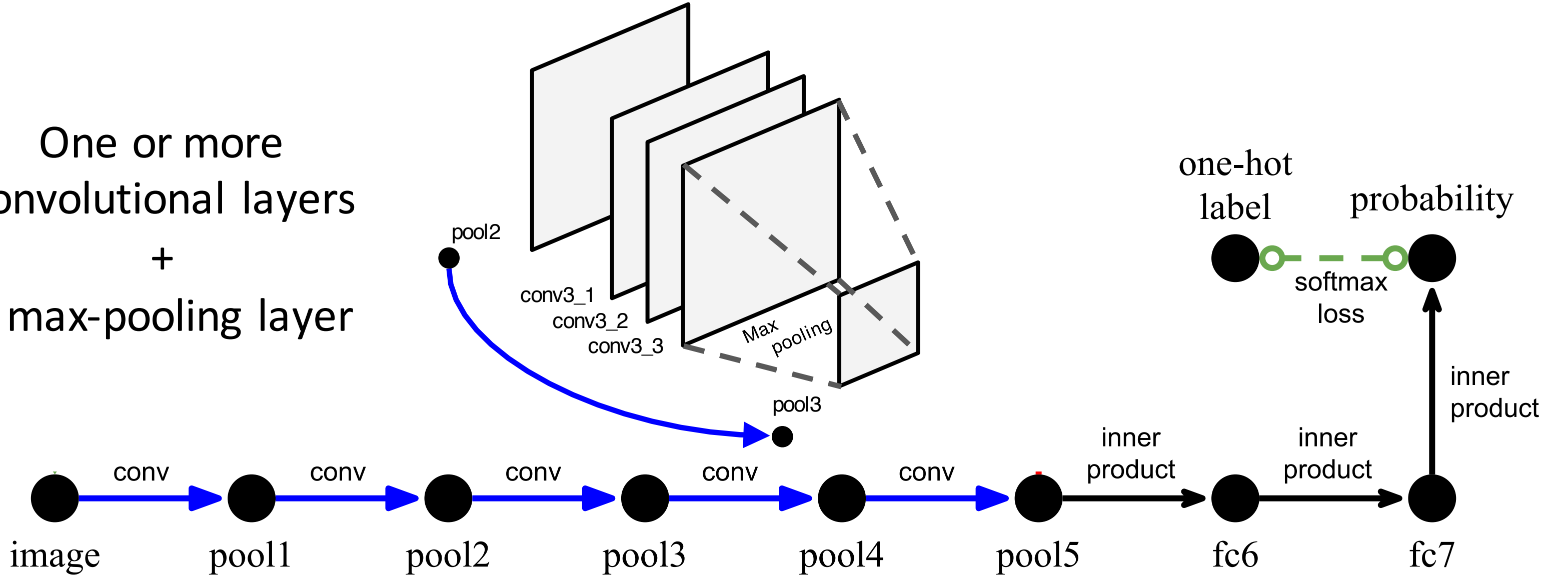
Outlines

1. The invertibility of large-scale image classification networks
2. Large-scale image classification networks with stronger invertibility

Invertibility of deep convolutional neural networks

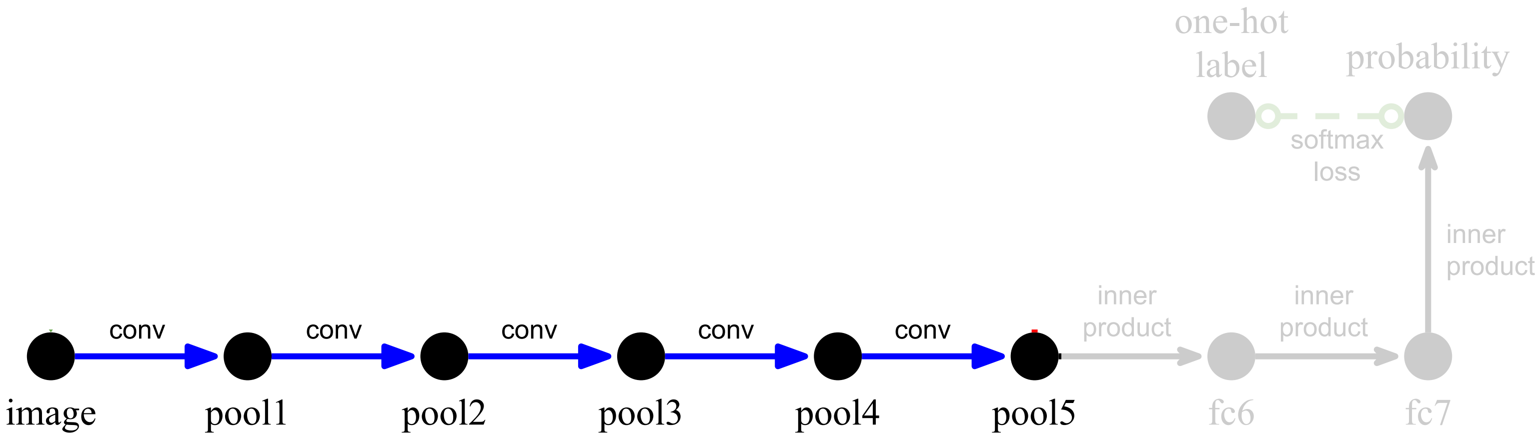
A typical classification network (VGGNet)

One or more convolutional layers
+
a max-pooling layer



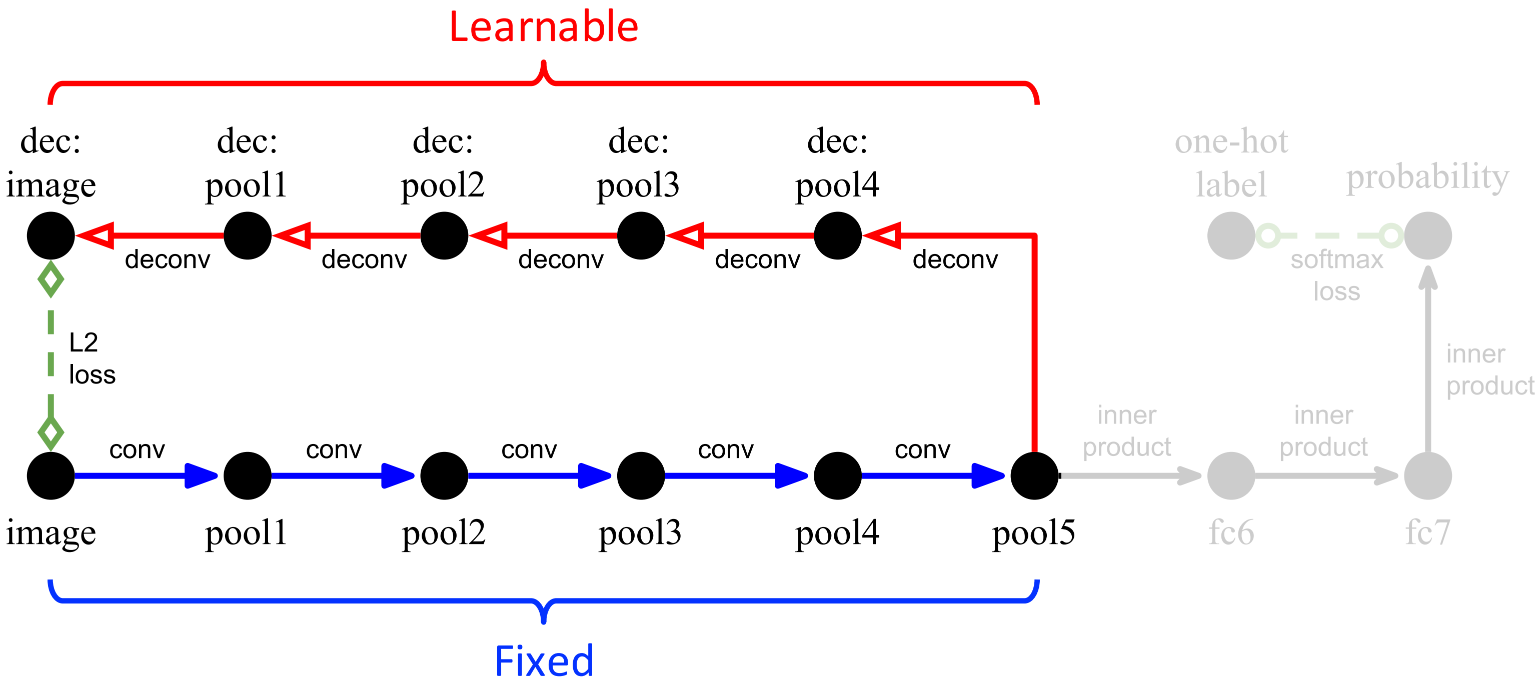
Yuting Zhang, Kibok Lee, Honglak Lee

Inducing an autoencoder from a classification network (VGGNet, pool5)



Yuting Zhang, Kibok Lee, Honglak Lee

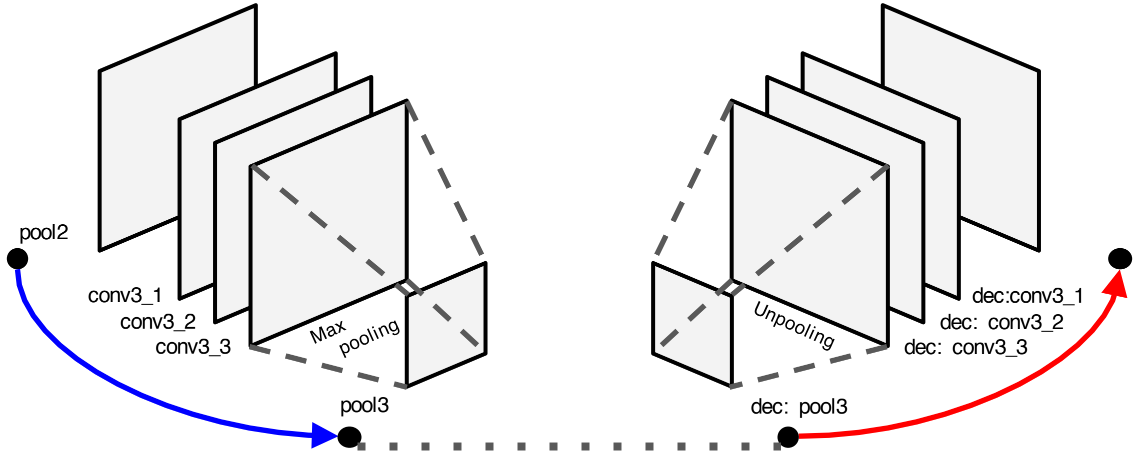
Training a decoding pathway for a classification network (VGGNet, pool5)



Yuting Zhang, Kibok Lee, Honglak Lee

Micro-architectures for decoders

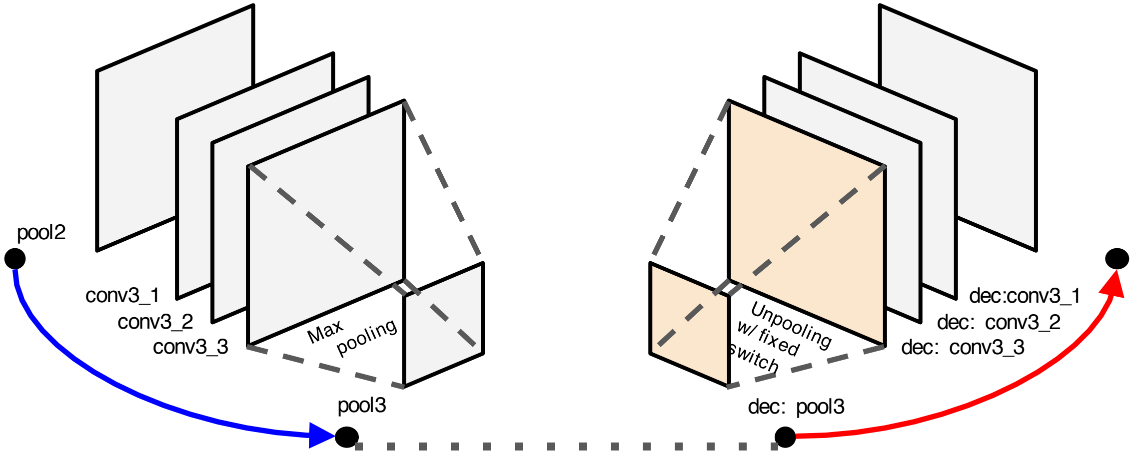
- Use “Unpooling” to approximately invert the pooling operation

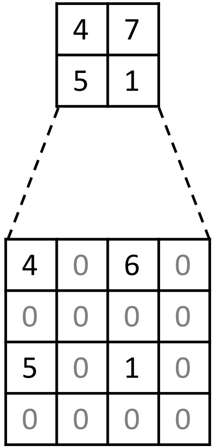


Yuting Zhang, Kibok Lee, Honglak Lee

Micro-architectures for decoders
(Unpooling with **fixed** switches, ordinary SAE)

- One can use the ordinary stacked autoencoder (SAE).
  - *Related work*: Dosovitskiy, A. and Brox, T, "Inverting visual representations with convolutional networks", CVPR 2016.



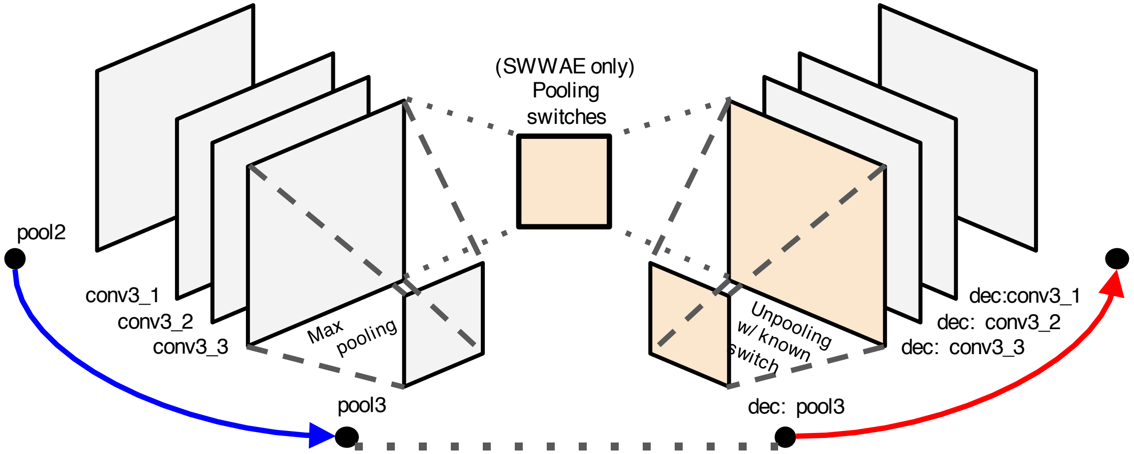


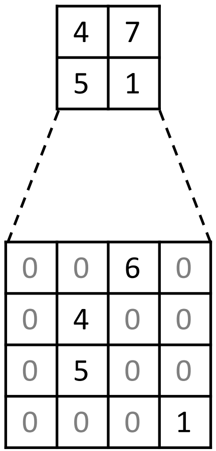
Unpooling with fixed switches (Upsampling)

Yuting Zhang, Kibok Lee, Honglak Lee

Micro-architectures for decoders
(Unpooling with **known** switches, SWWAE)

- We can also use stacked “what-where” autoencoders (SWWAE).
  - Unpooling with the known switches transferred from the encoder.
  - More accurate inversion, since spatial details are recovered better.





Unpooling with **known** switches

Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from **one layer**			

Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from pool1			

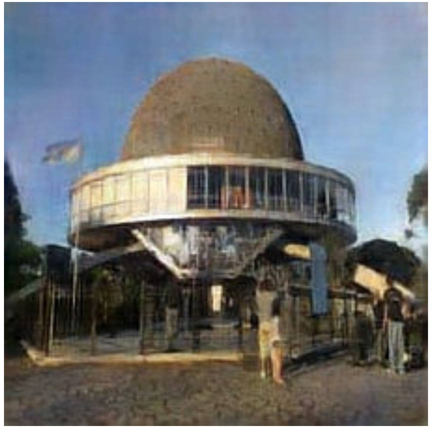
Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from **pool2**			

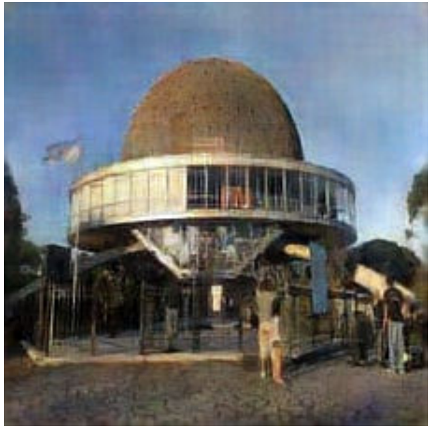
Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from pool3			

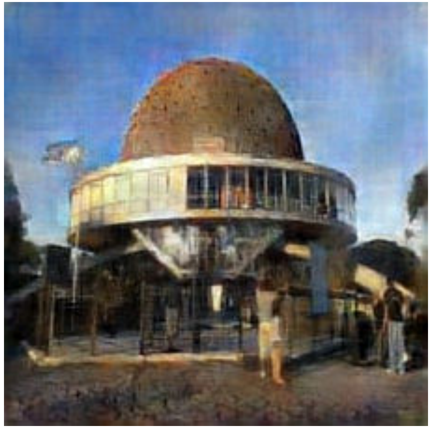
Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from **pool4**			

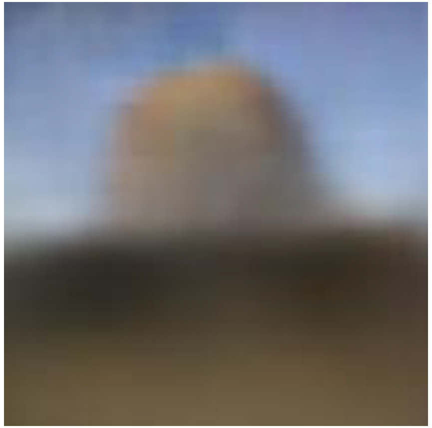
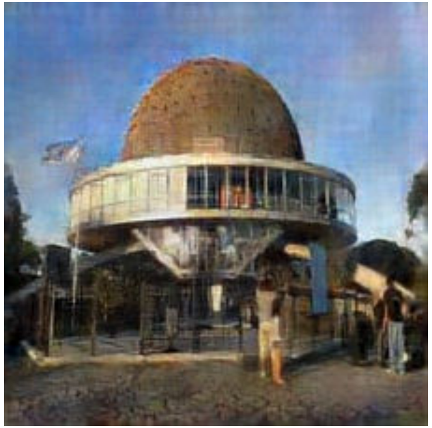
Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from pool5			

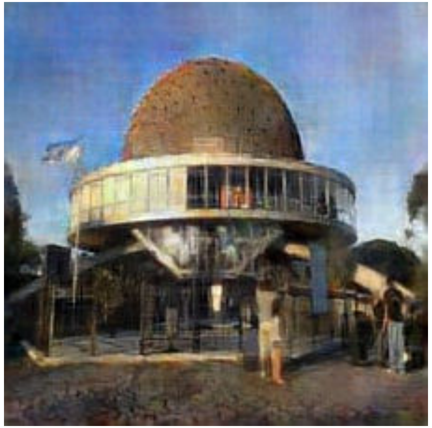
Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from fc6			

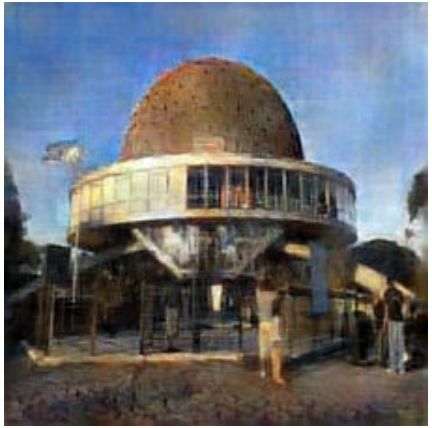
Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from fc7			

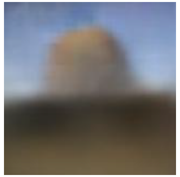
Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction from different layers (AlexNet)

	Input Image	SAE Dosovitskiy & Brox (2015)	SWWAE
Reconstructed from fc8			

Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction via SAE decoders

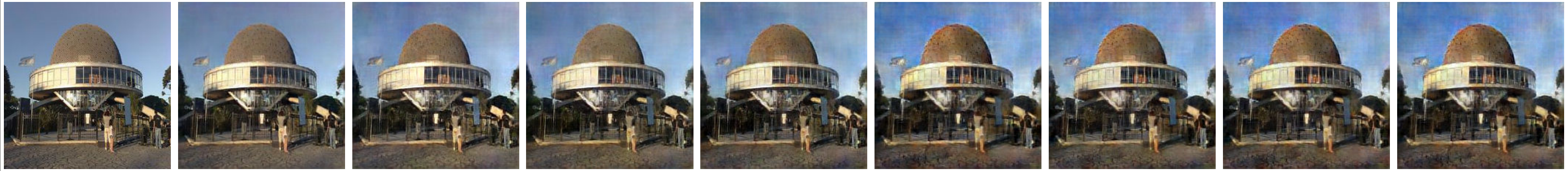
Layer	image	pool1	pool2	conv3	conv4	pool5	fc6	fc7	fc8
SAE Dosovitskiy & Brox (2016)									

The network is less invertible for higher layers,
so deeper representations preserve less input information.

- Two possible sources of information loss
  - Convolutional filters and non-linearity (Transformation)
  - Max-pooling (Spatial invariance)
- They are mixed in the SAE reconstruction results.

Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction via SAE and SWWAE decoders

Layer	image	pool1	pool2	conv3	conv4	pool5	fc6	fc7	fc8
SAE Dosovitskiy & Brox (2016)									
SWWAE									

- Using the encoder pooling switches for unpooling, the information loss due to max-pooling can be better recovered.
- The extremely good reconstruction quality of SWWAE indicates the "convolutional filters + ReLU" cause **very minor** information losses.

Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction for 16-layer **VGGNet**

Layer	image	pool1	pool2	pool3	pool4	pool5
SAE						
SWWAE						

Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction for 16-layer **VGGNet**

Layer	image	pool1	pool2	pool3	pool4	pool5
SAE						
SWWAE						

Yuting Zhang, Kibok Lee, Honglak Lee

Reconstruction for 16-layer **VGGNet**

Layer	image	pool1	pool2	pool3	pool4	pool5
SAE						
SWWAE						

Yuting Zhang, Kibok Lee, Honglak Lee

Observations from reconstruction

Operator	Effect	Information loss
Convolutional filters + ReLU	Feature transformation	Minor
Max-pooling	Spatial invariance	Significant

Hypotheses from reconstruction

Operator	Effect	Information loss
Convolutional filters + ReLU	Feature transformation	**The less, the better**

- The invertiblility is important and potentially helpful for the convolutional filters in a deep classification network.

Yuting Zhang, Kibok Lee, Honglak Lee

Hypotheses from reconstruction

Operator	Effect	Information loss
Convolutional filters + ReLU	Feature transformation	**The less, the better**

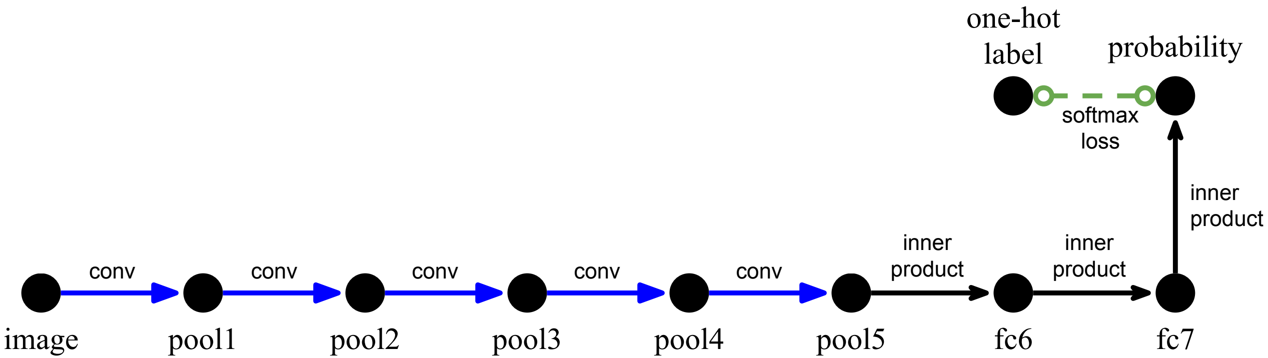
Aim to improve the classification network

Yuting Zhang, Kibok Lee, Honglak Lee

Classification networks with stronger invertibility

Classification networks with stronger invertibility

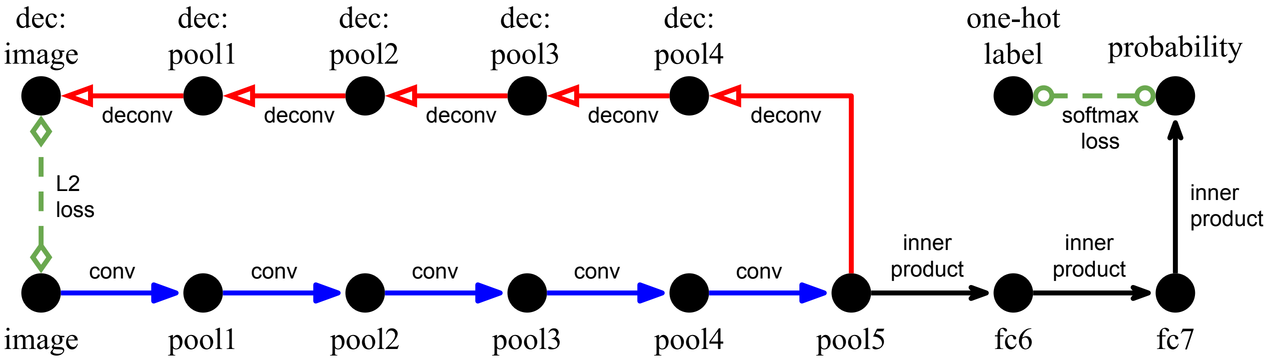
- Given a classification network
  - We take the 16-layer VGGNet as the baseline model



Yuting Zhang, Kibok Lee, Honglak Lee

Classification networks with stronger invertibility

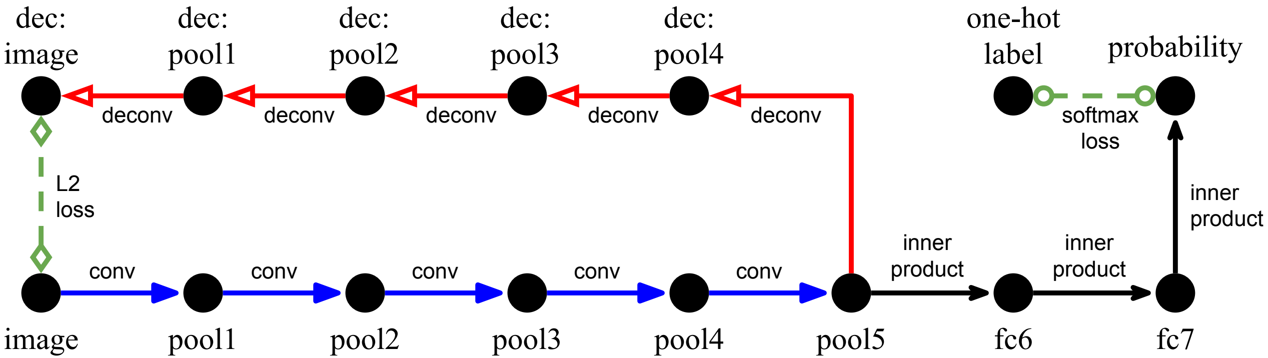
- Augmenting the classification network with a decoding pathway
  - starting from the last convolutional layer (pool5 in VGGNet)
- Multi-task learning using both classification and reconstruction objectives.



Yuting Zhang, Kibok Lee, Honglak Lee

Training procedure

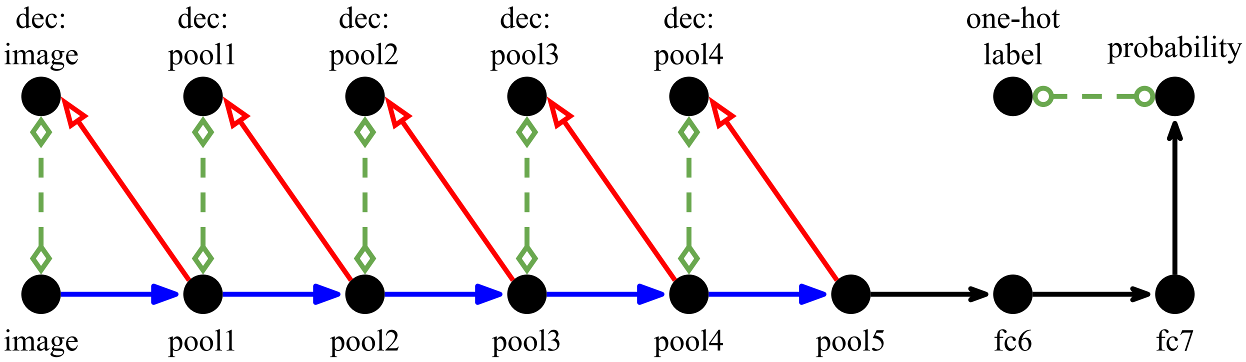
- *Step 1*: Initialize the classification network with pretrained weights.
- *Step 2*: Train the decoder while fixing the classification network.
  - For very deep network, it is hard to train it directly with random initialization.



Yuting Zhang, Kibok Lee, Honglak Lee

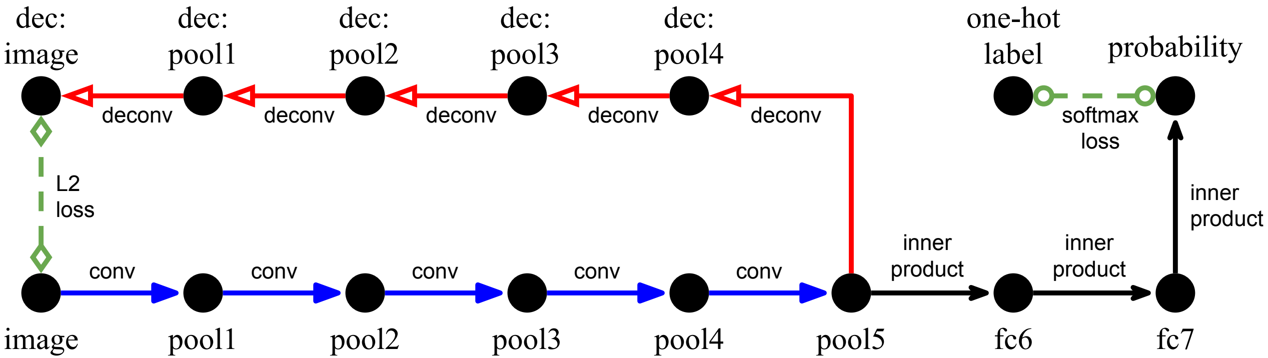
Model variant: SAE/SWWAE-layerwise

- *Step 2*: Train "layerwise" decoding pathways from random initialization.



Model variant: SAE/SWWAE-first

- *Step 2*: Train "layerwise" decoding pathways from random initialization.
- *Step 3*: Train the top-down decoding pathways, which is initialized in Step 2.
  - The reconstruction loss is only at the "first" layer.

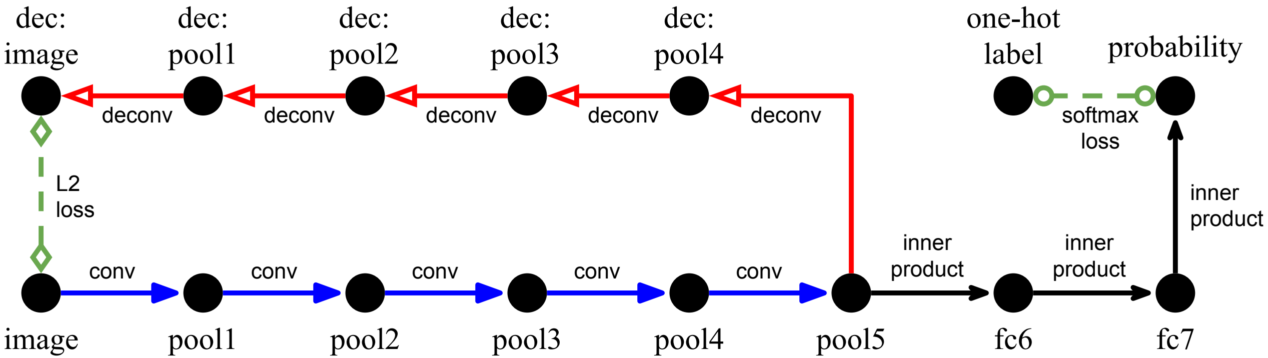


Yuting Zhang, Kibok Lee, Honglak Lee

Training procedure

- *Step 2*: Train "layerwise" decoding pathways from random initialization.
- *Step 3*: Train the top-down decoding pathways, which is initialized in Step 2.
- *Step 4*: Finetune the entire augmented network together.

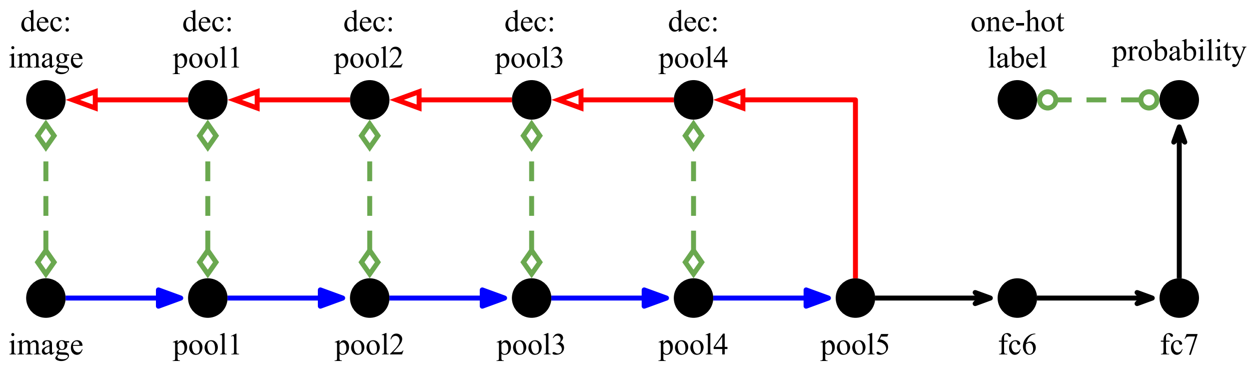
Mini-batch SGD for all steps



Yuting Zhang, Kibok Lee, Honglak Lee

Model variant: SAE/SWWAE-all

- Every layer can have its own reconstruction loss
  - Decoder layers can better corresponds to encoder layers
  - Intermediate layers can get more training signals



Yuting Zhang, Kibok Lee, Honglak Lee

Evaluations on ImageNet ILSVRC 2012

- Baseline model: 16-Layer VGGNet
- Augmented models: SAE/SWWAE - first/all/layerwise (6 in total)
- Testing protocol
  - Rescaling the shorter edge to 256px
  - **“Single crop”** scheme: 224x224 patch in the center
    - Clean results without postprocessing
  - **“Convolution”** scheme: whole VGGNet as a convolutional operator
    - More practical results

Validation errors on ImageNet ILSVRC 2012

Sampling	Single crop	
Model	Top-1	Top-5
VGGNet	29.05	10.07

Yuting Zhang, Kibok Lee, Honglak Lee

Validation errors on ImageNet ILSVRC 2012

Sampling	Single crop	
Model	Top-1	Top-5
VGGNet	29.05	10.07
+ SAE-first	27.70	9.28
+ SAE-all	**27.54**	**9.17**
+ SAE-layerwise	27.60	9.19

Get lower errors

Validation errors on ImageNet ILSVRC 2012

Sampling	Single crop	
Model	Top-1	Top-5
VGGNet	29.05	10.07
+ SAE-first	27.70	9.28
+ SAE-all	**27.54**	**9.17**
+ SAE-layerwise	27.60	9.19

Layer-wise reconstruction loss is helpful.

Yuting Zhang, Kibok Lee, Honglak Lee

Validation errors on ImageNet ILSVRC 2012

Sampling	Single crop	
Model	Top-1	Top-5
VGGNet	29.05	10.07
+ SAE-first	27.70	9.28
+ SAE-all	27.54	9.17
+ SAE-layerwise	27.60	9.19
+ SWWAE-first	27.60	9.23
+ SWWAE-all	**27.39**	**9.06**
+ SWWAE-layerwise	27.53	9.10

Layer-wise reconstruction loss is helpful.

Even lower errors

Yuting Zhang, Kibok Lee, Honglak Lee

Validation errors on ImageNet ILSVRC 2012

Sampling	Single crop	
Model	Top-1	Top-5
VGGNet	29.05	10.07
+ SAE-first	27.70	9.28
+ SAE-all	27.54	9.17
+ SAE-layerwise	27.60	9.19
+ SWWAE-first	27.60	9.23
+ SWWAE-all	**27.39**	**9.06**
+ SWWAE-layerwise	27.53	9.10

Layer-wise reconstruction loss is helpful.

SWWAE performs slightly better than ordinary SAE

Yuting Zhang, Kibok Lee, Honglak Lee

Validation errors on ImageNet ILSVRC 2012

Sampling	Single crop		Convolution	
Model	Top-1	Top-5	Top-1	Top-5
VGGNet	29.05	10.07	26.97	8.94
+ SAE-first	27.70	9.28	26.09	8.30
+ SAE-all	27.54	9.17	26.10	8.21
+ SAE-layerwise	27.60	9.19	26.06	8.17
+ SWWAE-first	27.60	9.23	25.87	8.14
+ SWWAE-all	**27.39**	**9.06**	**25.79**	**8.13**
+ SWWAE-layerwise	27.53	9.10	25.97	8.20

Training errors on ImageNet ILSVRC 2012

Sampling	Single crop	
Model	Top-1	Top-5
VGGNet	17.43	4.02
+ SAE-first	15.36	3.13
+ SAE-all	15.64	3.23
+ SAE-layerwise	16.20	3.42
+ SWWAE-first	15.10	3.08
+ SWWAE-all	15.67	3.24
+ SWWAE-layerwise	15.42	3.32

Training errors on ImageNet ILSVRC 2012

Sampling	Single crop	
Model	Top-1	Top-5
VGGNet	17.43	4.02
+ SAE-first	15.36	3.13
+ SAE-all	15.64	3.23
+ SAE-layerwise	16.20	3.42
+ SWWAE-first	15.10	3.08
+ SWWAE-all	15.67	3.24
+ SWWAE-layerwise	15.42	3.32

Get lower training errors

The unsupervised objectives help with the optimization of the supervised objectives.

Yuting Zhang, Kibok Lee, Honglak Lee

Training errors on ImageNet ILSVRC 2012

Sampling	Single crop		Validation errors	
Model	Top-1	Top-5	Top-1	Top-5
+ SAE-first	**15.36**	**3.13**	26.09	8.30
+ SAE-all	15.64	3.23	**26.10**	**8.21**
+ SWWAE-first	**15.10**	**3.08**	25.87	8.14
+ SWWAE-all	15.67	3.24	**25.79**	**8.13**

Yuting Zhang, Kibok Lee, Honglak Lee

Training errors on ImageNet ILSVRC 2012

Sampling	Single crop		Validation errors	
Model	Top-1	Top-5	Top-1	Top-5
+ SAE-first	**15.36**	**3.13**	26.09	8.30
+ SAE-all	15.64	3.23	**26.10**	**8.21**
+ SWWAE-first	**15.10**	**3.08**	25.87	8.14
+ SWWAE-all	15.67	3.24	**25.79**	**8.13**

Compared to SAE/SWWAE-first, SAE/SWWAE-all has

- higher training errors
- lower validation errors

Layer-wise reconstruction loss has regularization effects.

Yuting Zhang, Kibok Lee, Honglak Lee

Conclusions

- A simple and effective way to incorporate unsupervised objectives into large-scale classification network learning.
- The resultant autoencoder can reconstruct image with extremely high quality from deep representations.
- We improved the image classification performance of the 16-layer VGGNet, a strong baseline model, by a noticeable margin.
- We hope this paper will inspire further investigations on the use of unsupervised learning in a large-scale setting.

Thank you!

Full version: arxiv.org/abs/1606.06582

Code (GitHub): bit.ly/cnn-dec