

CVPR 2014 Tutorial

Deep Learning for Computer Vision

Multimodal learning and Multitask learning

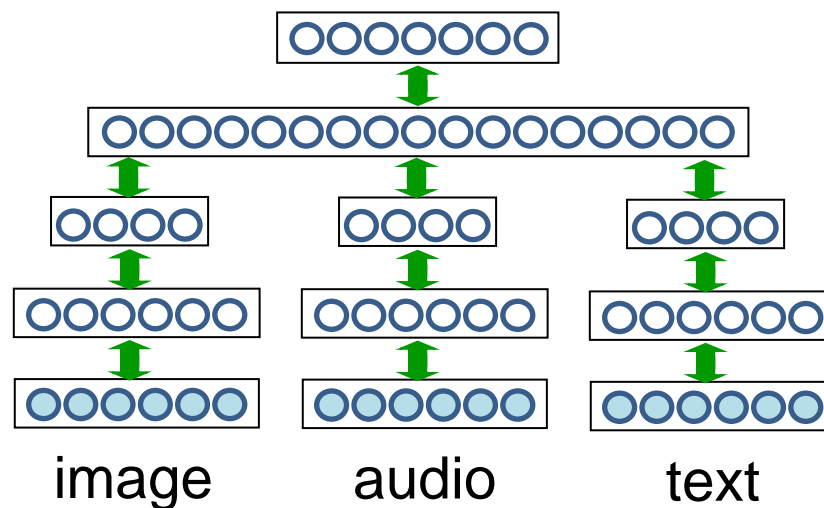
Honglak Lee (University of Michigan)

Outline

- Multimodal Deep Learning
 - Audio + Video
 - Image + Text
- Deep Transfer/Multitask Learning
 - Generalization of deep learning features over multi tasks
 - Disentangling factors of variations

Multimodal Deep Learning

- Motivation: Single deep learning algorithms that combine multiple input domains
 - Images
 - Audio & speech
 - Video
 - Text
 - Robotic sensors
 - Time-series data
 - ...



Multimodal Deep Learning

- Advantages
 - Improved recognition performance
 - Robustness to missing values or missing modalities
- Key questions
 - How can we capture associations between heterogeneous modalities by learning mid-level feature representations?

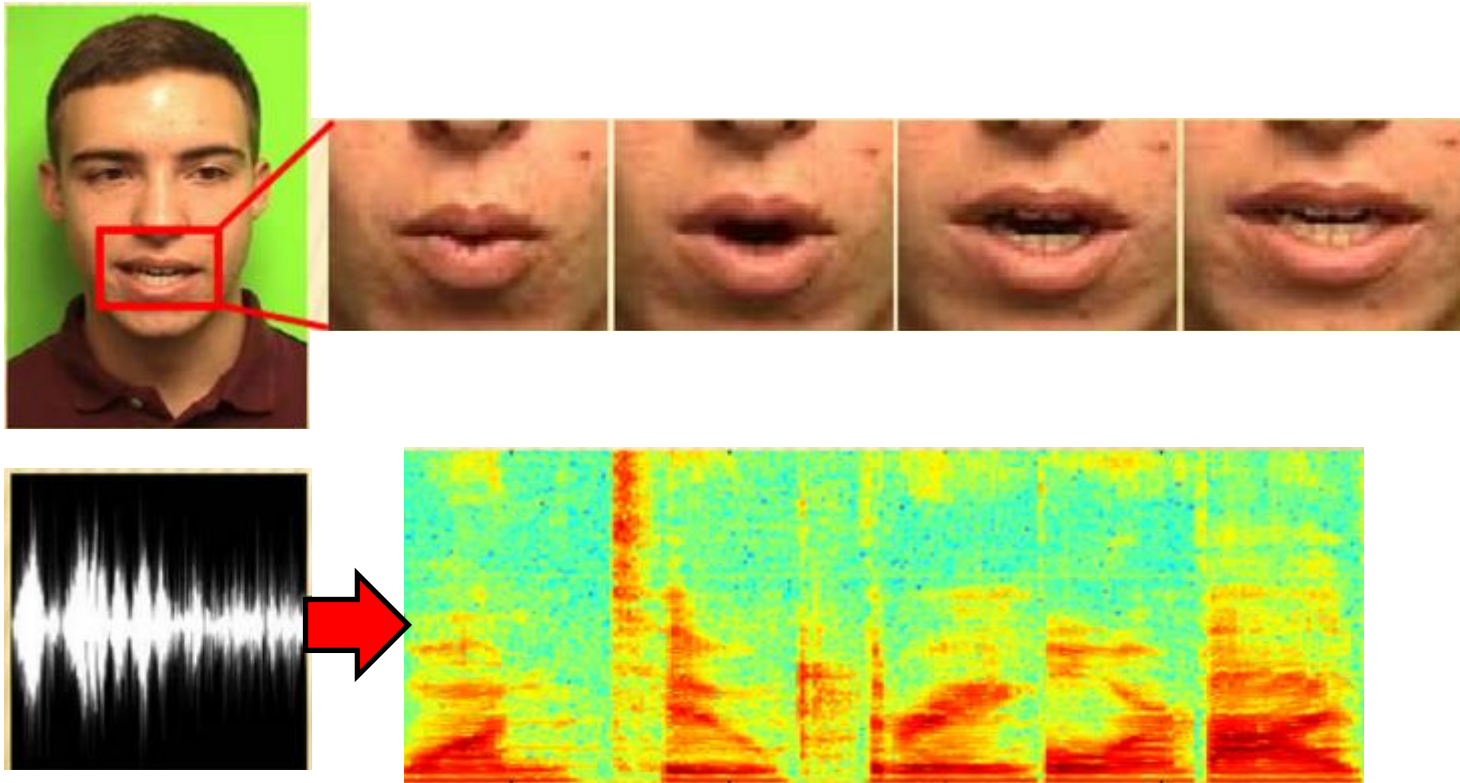
Multimodal deep learning from audio-visual data

Audio visual speech recognition

- Key problems
 - Can we improve “lip-reading” performance by learning features from video and speech?
 - Does multimodal feature learning improve speech recognition?
 - Can we learn shared representation that can do robust recognition when some modalities are missing at test time?
- Related work
 - Potamianos et al., Audio-visual automatic speech recognition: An overview, Issues in Visual and Audio-Visual Speech Processing. 2004
 - Matthews et al., Extraction of visual features for lipreading, PAMI 2004
 - Gurban, M. and Thiran, J.P. Information theoretic feature extraction for audio-visual speech recognition. IEEE TSP, 2009

Multimodal Feature Learning

- Lip reading via multimodal feature learning (audio / visual data)

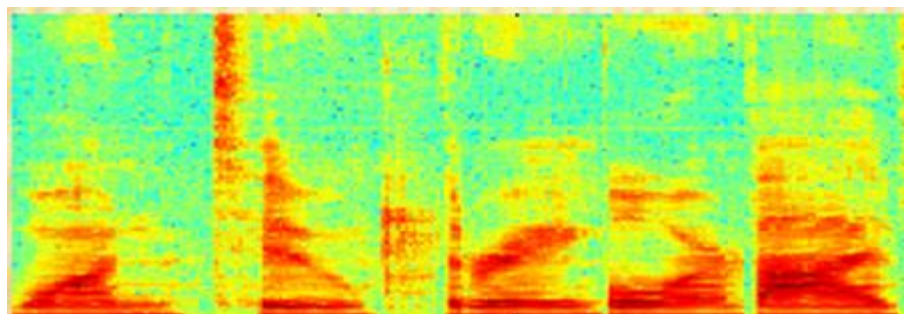


Multimodal Feature Learning

- Lip reading via multimodal feature learning (audio / visual data)



$$\begin{bmatrix} 1.5 \\ -0.1 \\ 0 \\ 0.3 \\ \cdot \\ \cdot \\ \cdot \\ 2.0 \end{bmatrix}$$

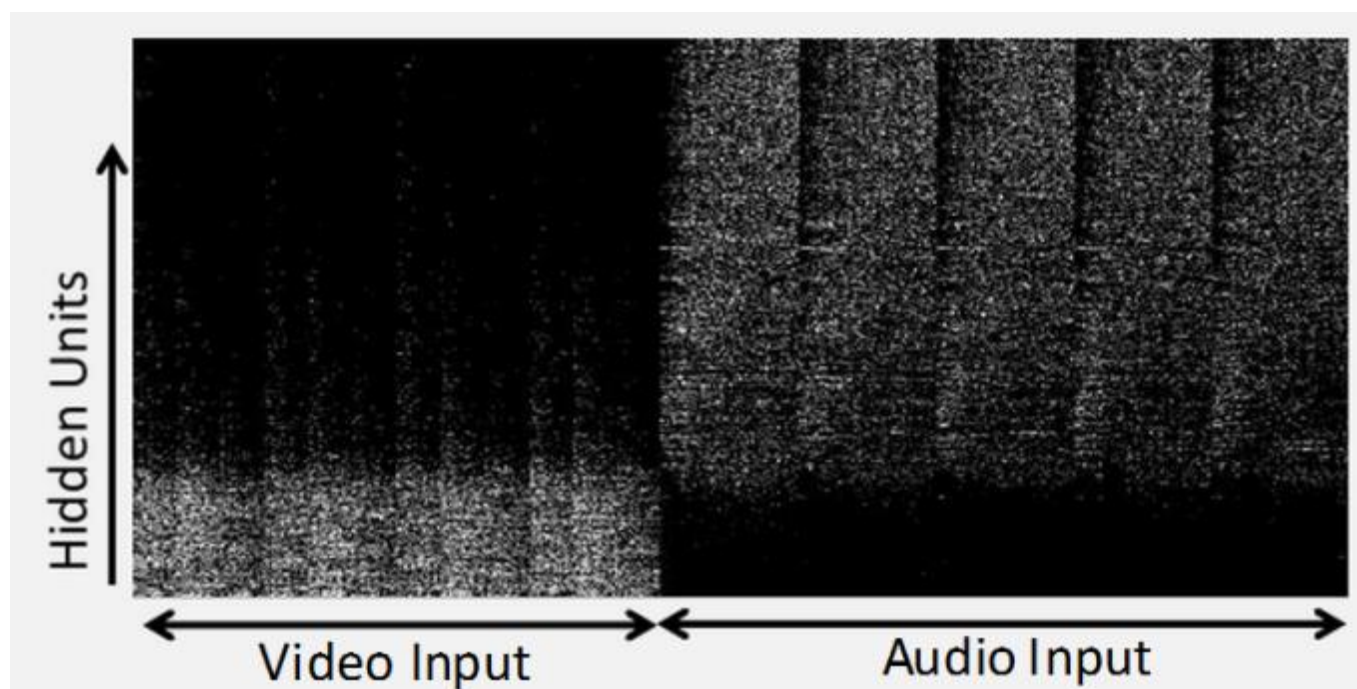


$$\begin{bmatrix} 0 \\ -0.3 \\ 0 \\ -0.8 \\ \cdot \\ \cdot \\ \cdot \\ 1.1 \end{bmatrix}$$

Q. Is concatenating the best option?

Multimodal Feature Learning

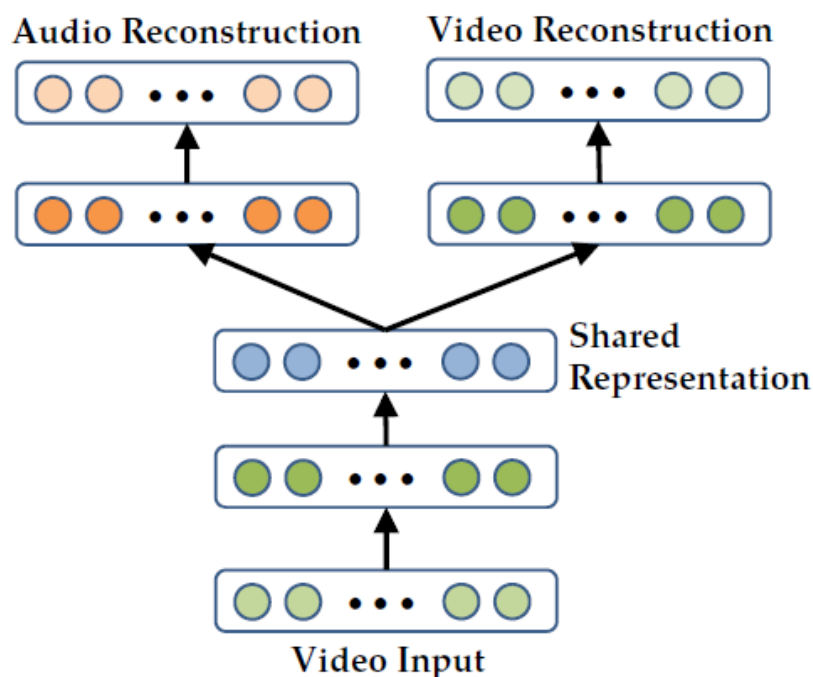
- Concatenating and learning features (via a single layer learning) doesn't work



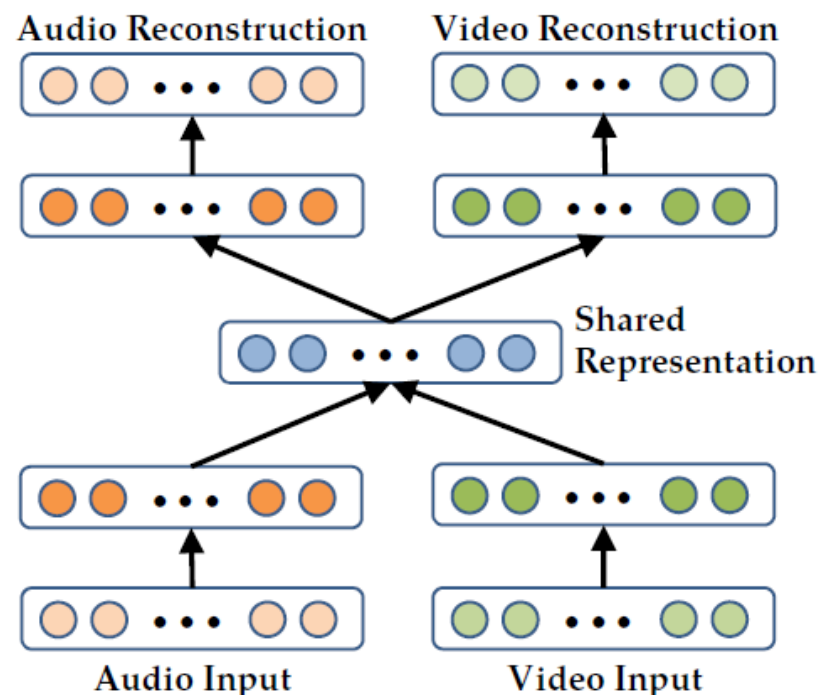
Mostly “unimodal” features are learned

Multimodal Feature Learning

- Bimodal autoencoder
 - Idea: predict unseen modality from observed modality



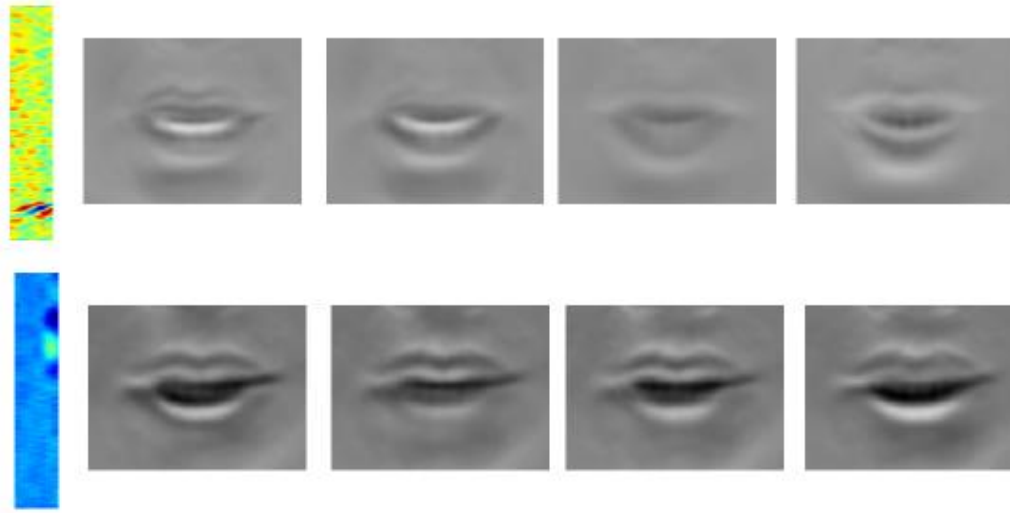
(a) Video-Only Deep Autoencoder



(b) Bimodal Deep Autoencoder

Multimodal Feature Learning

- Visualization of learned filters



Audio(spectrogram) and Video features learned over 100ms windows

- Results: AVLetters Lip reading dataset

Method	Accuracy
Zhao et al. (IEEE Multimedia 2009)	58.9%
Multimodal deep autoencoder (Ngiam et al., ICML 2011)	65.8%

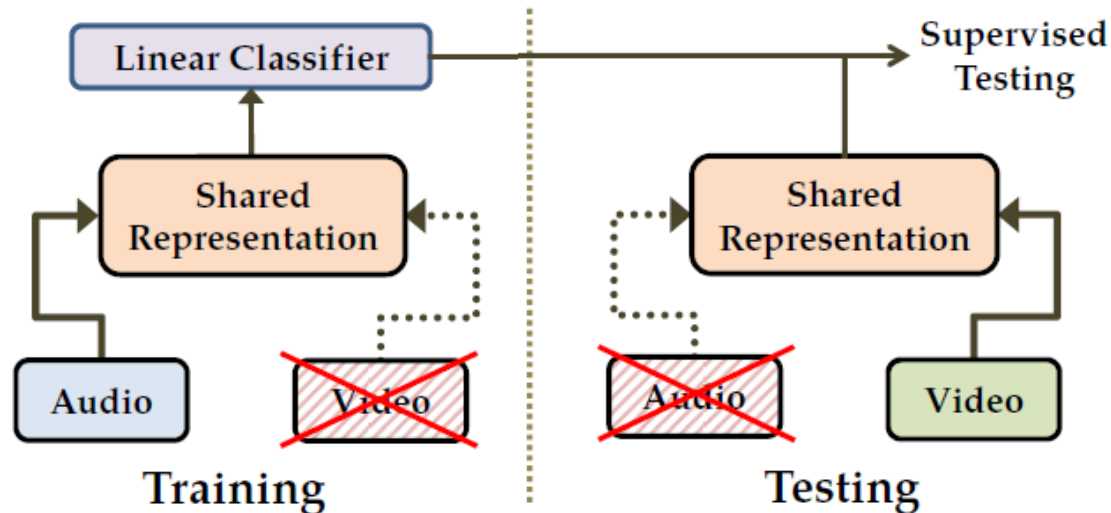
Speech recognition with noise

- Multimodal feature learning improves speech recognition when the audio input is corrupted with noise
 - Classification on CUAVE dataset (digit recognition)

Feature Representation	Accuracy (Clean Audio)	Accuracy (Noisy Audio)
Audio RBM	95.8%	75.8% \pm 2.0%
Multimodal DAE	90.0%	77.3% \pm 1.4%
Multimodal DAE+ Audio RBM	94.4%	82.2% \pm 1.2%

Robustness to missing modalities

- “Learning to see” experiments



- Performance on CUAVE dataset

Train / Test	Method	Accuracy
Audio -> Video	Raw-CCA features	41.9%
	“deep”-CCA features	57.3%
Video -> Audio	Raw-CCA features	42.9%
	“deep”-CCA features	91.7%

Multimodal deep learning from image & text

Learning from images and text

- Key problems
 - Improving robustness given images and text
 - Generating text descriptors from images
 - Retrieving images from text queries
- Related work
 - M. Guillaumin, J. Verbeek, and C. Schmid, CVPR 2010
 - M. Huiskes, B. Thomee, and M. Lew, Multimedia Information Retrieval, 2010
 - E. Xing, R. Yan, and A. Hauptmann. UAI 2005
 - G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, CVPR 2011
 - ...

Training Data

- Samples from the MIR Flickr Dataset



pentax, k10d,
kangarooisland
southaustralia, sa
australia
sealion 300mm



camera, jahdakine,
lightpainting,
reflection
doublepaneglass
wowiekazowie



sandbanks, lake,
lakeontario, sunset,
walking, beach,
purple, sky,
water, clouds,
overtheexcellence



top20buperflies



<no text>



mickikrimmel,
mickipedia, headshot

Tasks

- Improve Classification



pentax, k10d, kangarooisland
southaustralia, sa australia
australiansealion 300mm



SEA / NOT SEA

- Fill in Missing Modalities (image -> text)



beach, sea, surf,
strand, shore,
wave, seascape,
sand, ocean, waves

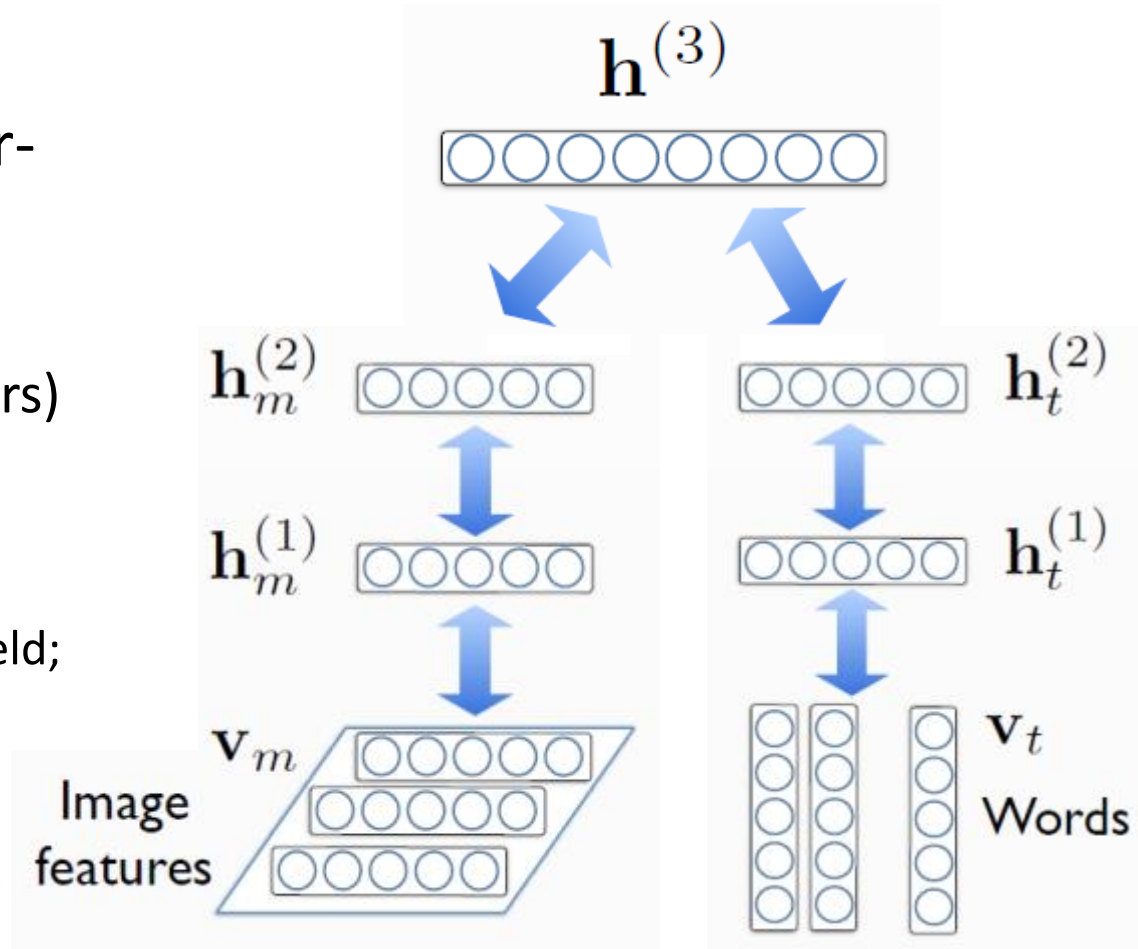
- Retrieve data from one modality when queried using data from another modality

beach, sea, surf,
strand, shore,
wave, seascape,
sand, ocean, waves



Multimodal Deep Boltzmann Machine

- Joint density model
- Modality specific lower-layers of DBM*
- Fusion at the top
(undirected across all layers)
- Inference:
 - Gibbs sampling or mean-field;
(bottom-up and top-down)
- Learning:
 - stochastic approximation











*R. Salakhutdinov & G. Hinton, Deep Boltzmann machine, AISTATS 2009

Inferring missing modalities

Generating tags given image

Image	Given Tags	Generated Tags
	pentax, k10d, kangarooisland, southaustralia, sa, australia, australiansealion, 300mm	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves
	<no text>	night, lights, christmas, nightshot, nacht, nuit, notte, longexposure, noche, nocturna
	aheram, 0505 sarahc, moo	portrait, bw, blackandwhite, woman, people, faces, girl, blackwhite, person, man
	unseulpixel, naturey crap	fall, autumn, trees, leaves, foliage, forest, woods, branches, path

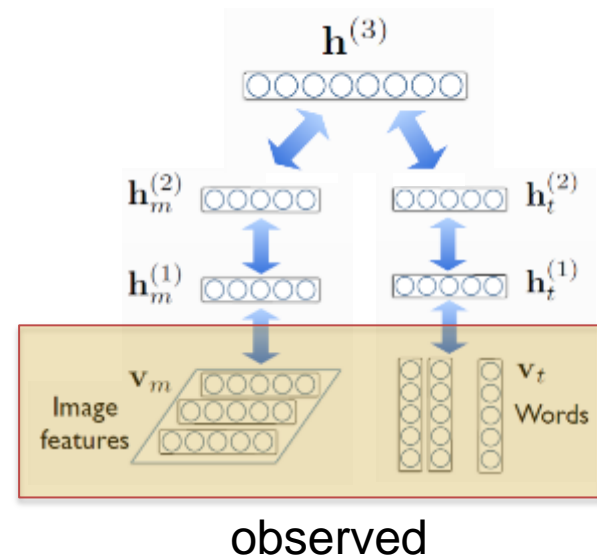
Retrieving images from text

Input Text	2 nearest neighbours to generated image features	
nature, hill scenery, green clouds		
flower, nature, green, flowers, petal, petals, bud		
blue, red, art, artwork, painted, paint, artistic surreal, gallery bleu		
bw, blackandwhite, noiret blanc, bianco enero, blancoynegro		

Classification Performance

- Multimodal inputs

Model	Mean AP
Random	0.124
LDA (Huiskes et al.)	0.492
SVM (Huiskes et al.)	0.475
DBM	0.609



- Unimodal (image-only) inputs

Model	Mean AP
Image-SVM (Huiskes et al.)	0.375
Image-only DBM	0.469
DBM (zero-out text inputs)	0.522
DBM (generate text)	0.531



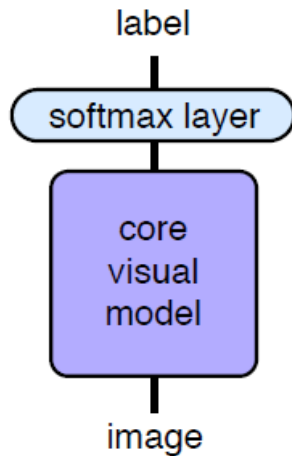
Deep Visual-Semantic Embedding

- Key idea
 - Two successful representations
 - Images: CNN features
 - Text (words): word embedding (via skipgram)
 - Associate image features and word embedding so that we can infer about “unknown” class
- Related work
 - H. Larochelle, D. Erhan, Y. Bengio. Zero-data Learning of New Tasks. AAAI 2008.
 - R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, A. Y. Ng. Zero-Shot Learning Through Cross-Modal Transfer. NIPS 2013.
 - M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, J. Dean. Zero-Shot Learning by Convex Combination of Semantic Embeddings. ArXiv 1312.5650.

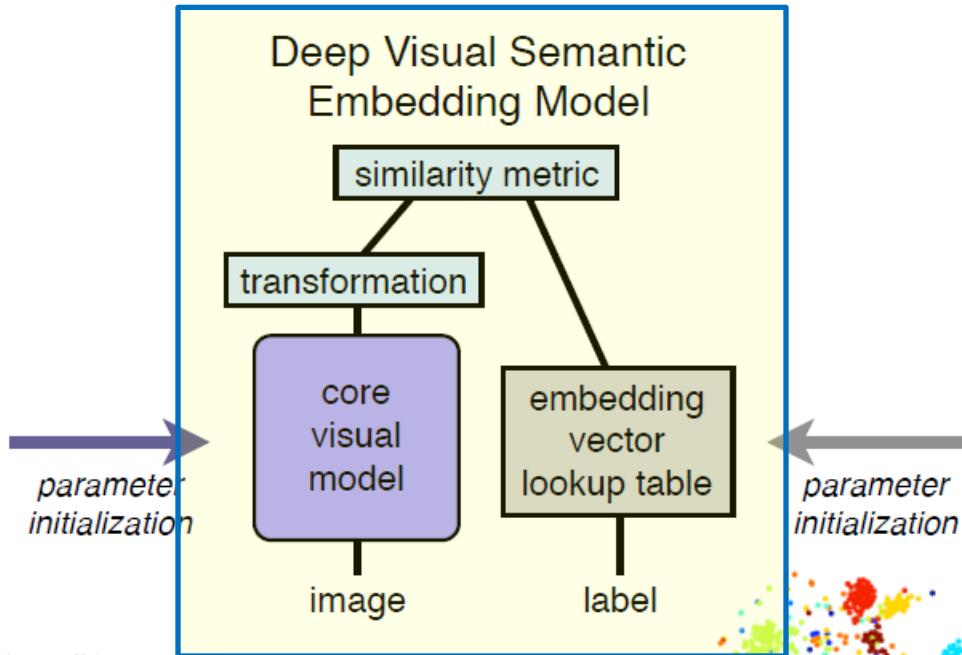
Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, DeViSE: A Deep Visual-Semantic Embedding Model, NIPS 2013

Deep Visual-Semantic Embedding

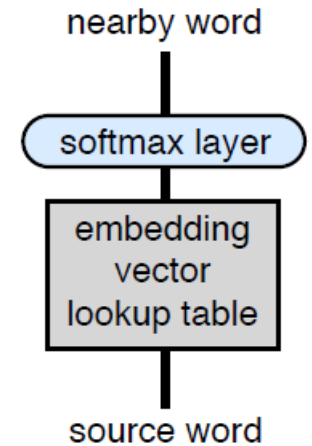
Traditional Visual Model



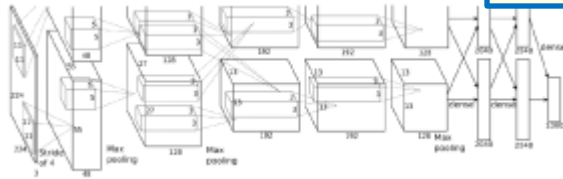
Deep Visual Semantic Embedding Model



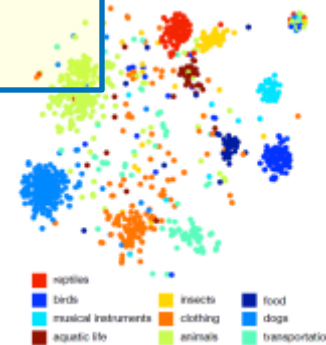
Skip-gram Language Model



CNN



Skipgram for word embedding



Visualization of label embedding

$$\text{loss}(\text{image}, \text{label}) = \sum_{j \neq \text{label}} \max[0, \text{margin} - \vec{t}_{\text{label}} M \vec{v}(\text{image}) + \vec{t}_j M \vec{v}(\text{image})]$$

Deep Visual-Semantic Embedding

- Zero-shot classification results
 - Train from Imagenet ILSVRC 1000 classes
 - Generalize to unseen classes

Data Set	Model	Hierarchical precision@ <i>k</i>				
		1	2	5	10	20
2-hop	DeViSE-0	0.06	0.152	0.192	0.217	0.233
	DeViSE+1K	0.008	0.204	0.196	0.201	0.214
	Softmax baseline	0	0.236	0.181	0.174	0.179
3-hop	DeViSE-0	0.017	0.037	0.191	0.214	0.236
	DeViSE+1K	0.005	0.053	0.192	0.201	0.214
	Softmax baseline	0	0.053	0.157	0.143	0.130
ImageNet 2011 21K	DeViSE-0	0.008	0.017	0.072	0.085	0.096
	DeViSE+1K	0.003	0.025	0.083	0.092	0.101
	Softmax baseline	0	0.023	0.071	0.069	0.065

more zero-shot
classes to
predict
(harder)



Summary: Multimodal deep learning

- Multimodal deep learning can construct shared representation that associates between heterogeneous modalities
 - Improved recognition performance
 - Robust to missing modalities
 - Improved zero-shot learning

Outline

- Multimodal Deep Learning
 - Audio + Video
 - Image + Text
- Deep Transfer/Multitask Learning
 - Generalization of deep learning features over multi tasks
 - Disentangling factors of variations

Generalizable Learning

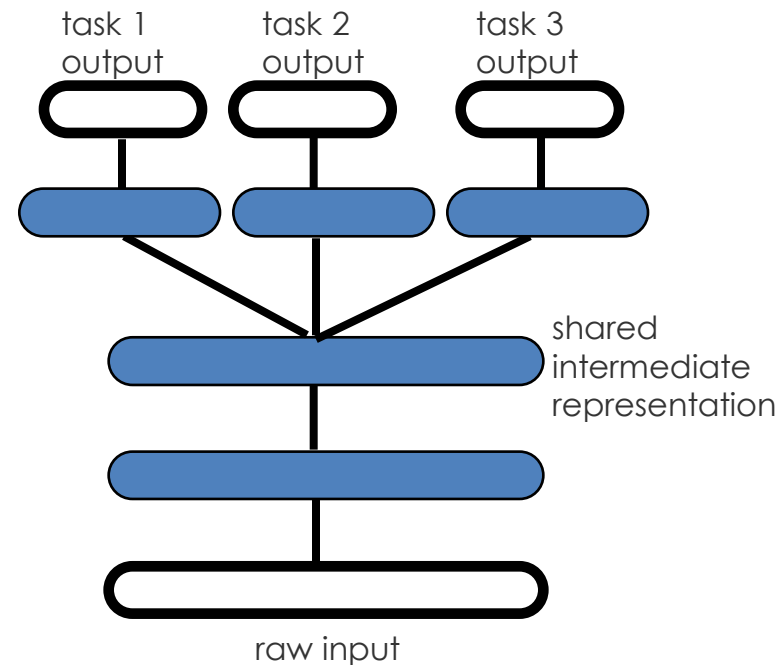
- Shared mid-level representations

- Multi-Task learning:

- Can we learn a shared representation that is useful for multiple tasks?

- Transfer learning:

- Can we learn a mid-level representation from a dataset that generalizes well to other datasets?



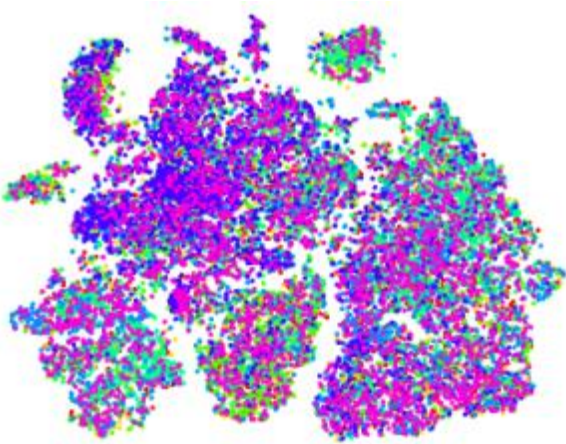
Related work: Caruana, 1999; Baxter, 2000, Thrun, 1996; Kumar, 2012; Evgeniou & Pontil 2004; Argiryou, 2007; Chen et al., 2011; Jacob et al., 2008

Slide credit: Y. Bengio

Generalization of deep learning
features over multi tasks

Feature Generalization

- Visualization of features (via t-SNE embedding)



Gist



DeCAF1



DeCAF6

ILSVRC-2012 validation set

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, ICML 2014

Feature Generalization

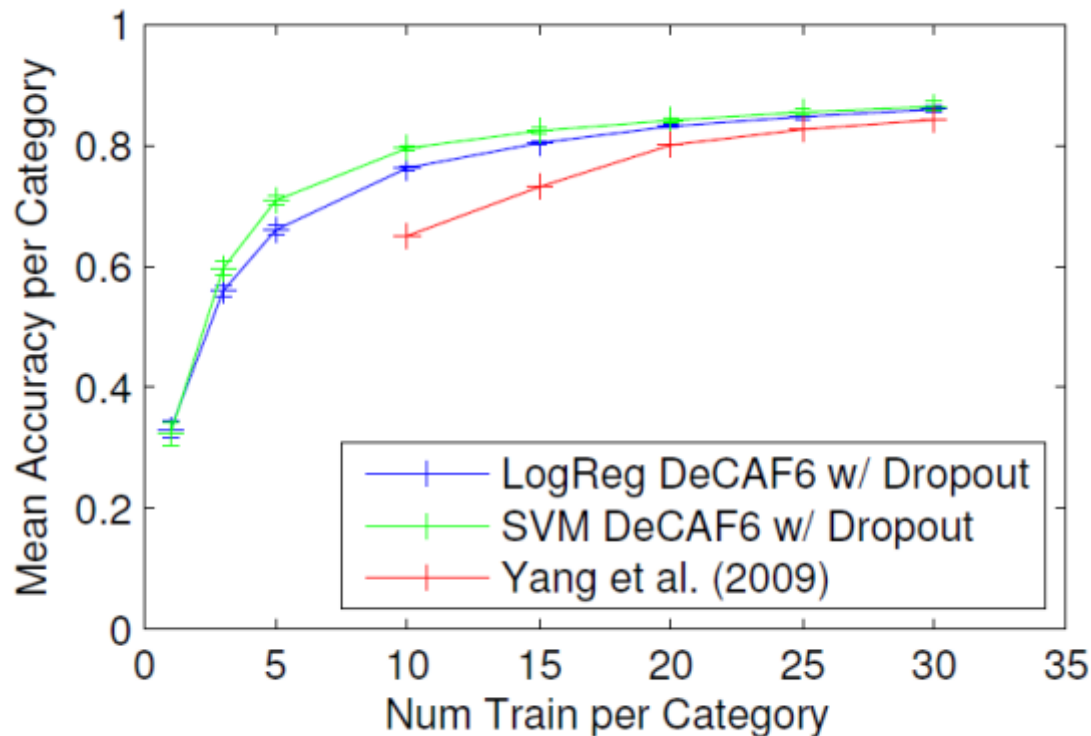
- Domain adaptation task

	Amazon \rightarrow Webcam		
	SURF	DeCAF ₆	DeCAF ₇
Logistic Reg. (S)	9.63 \pm 1.4	48.58 \pm 1.3	53.56 \pm 1.5
SVM (S)	11.05 \pm 2.3	52.22 \pm 1.7	53.90 \pm 2.2
Logistic Reg. (T)	24.33 \pm 2.1	72.56 \pm 2.1	74.19 \pm 2.8
SVM (T)	51.05 \pm 2.0	78.26 \pm 2.6	78.72 \pm 2.3
Logistic Reg. (ST)	19.89 \pm 1.7	75.30 \pm 2.0	76.32 \pm 2.0
SVM (ST)	23.19 \pm 3.5	80.66 \pm 2.3	79.12 \pm 2.1
Daume III (2007)	40.26 \pm 1.1	82.14 \pm 1.9	81.65 \pm 2.4
Hoffman et al. (2013)	37.66 \pm 2.2	80.06 \pm 2.7	80.37 \pm 2.0
Gong et al. (2012)	39.80 \pm 2.3	75.21 \pm 1.2	77.55 \pm 1.9
Chopra et al. (2013)		58.85	

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, ICML 2014

Feature Generalization

- Caltech 101 classification



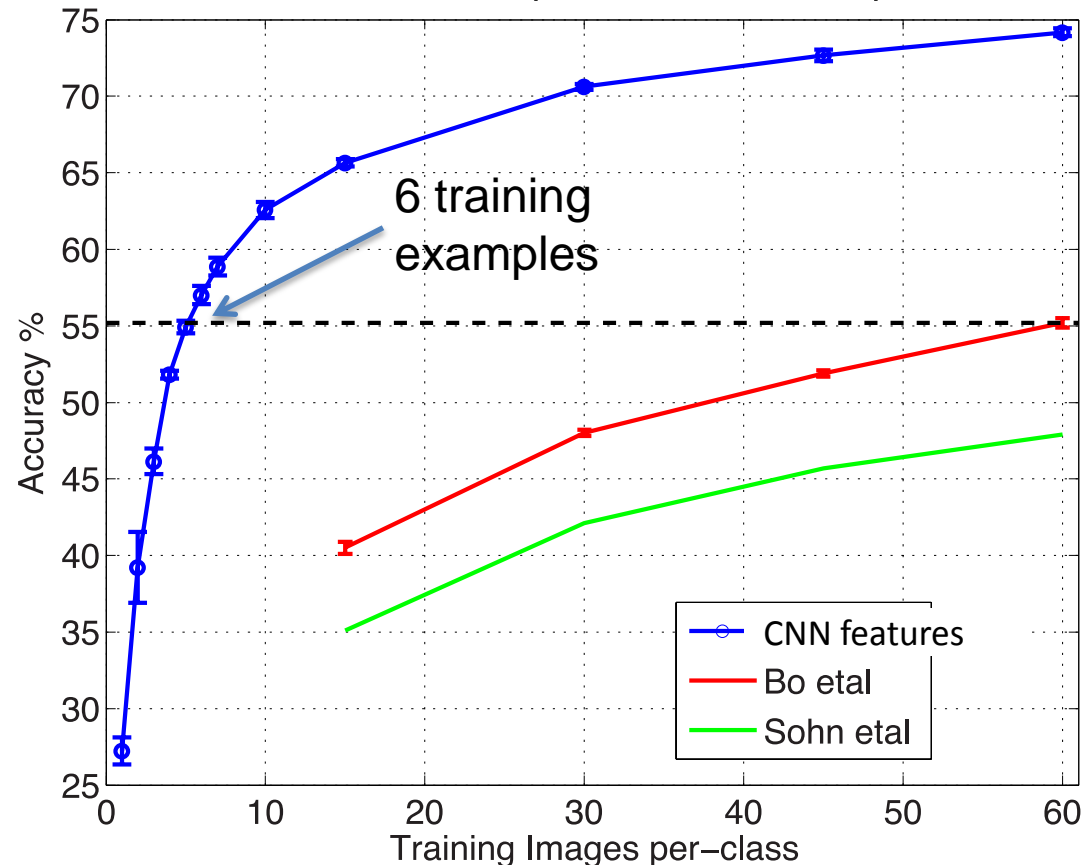
J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, ICML 2014

Feature Generalization

- Zeiler & Fergus, arXiv 1311.2901, 2013 (Caltech-101,256)
- Girshick et al. CVPR'14 (Caltech-101, SunS)
- Oquab et al. CVPR'14 (VOC 2012)
- Razavian et al. arXiv 1403.6382, 2014 (lots of datasets)

- Pre-train on
Imagnet

Retrain classifier
on Caltech256

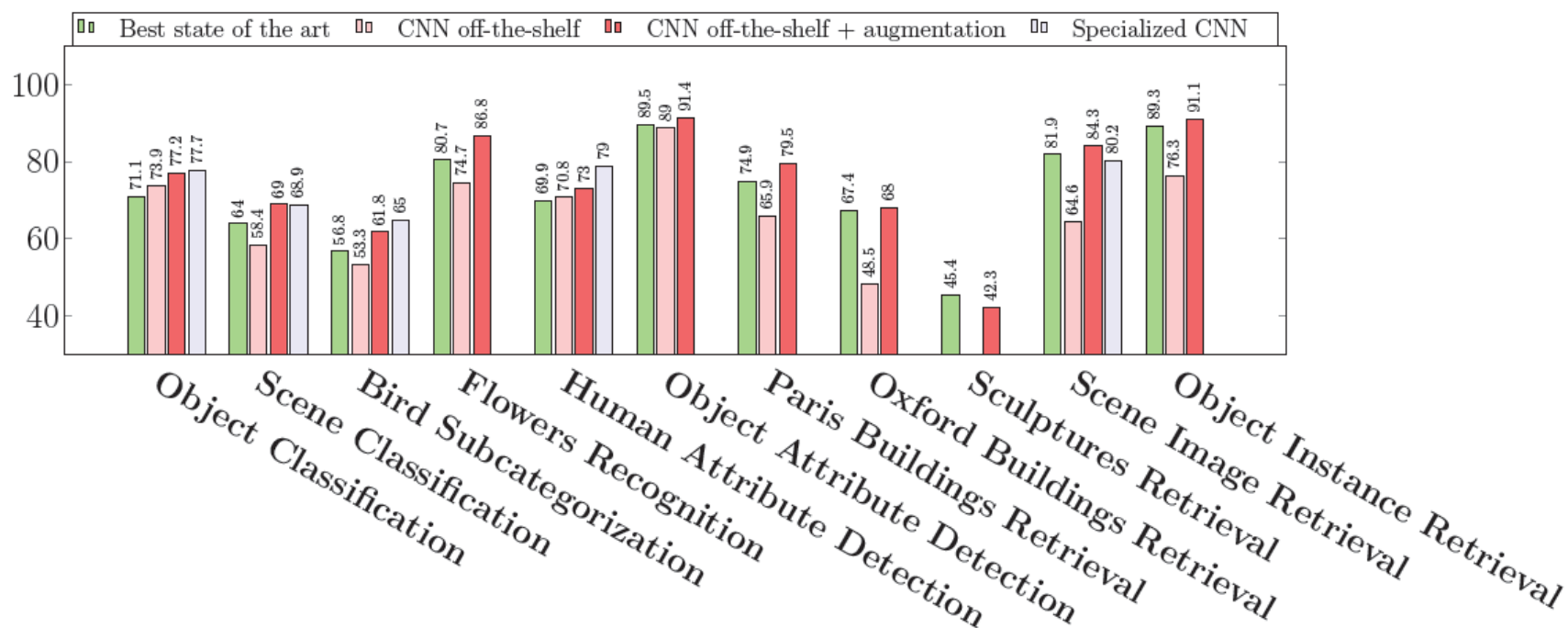


From Zeiler & Fergus, *Visualizing and Understanding Convolutional Networks*, arXiv 1311.2901, 2013

Slide credit: R. Fergus

Feature generalization over multiple tasks

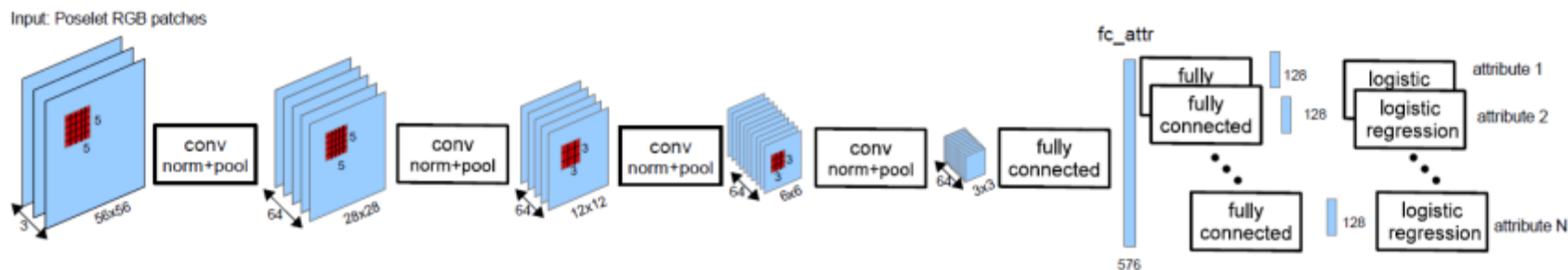
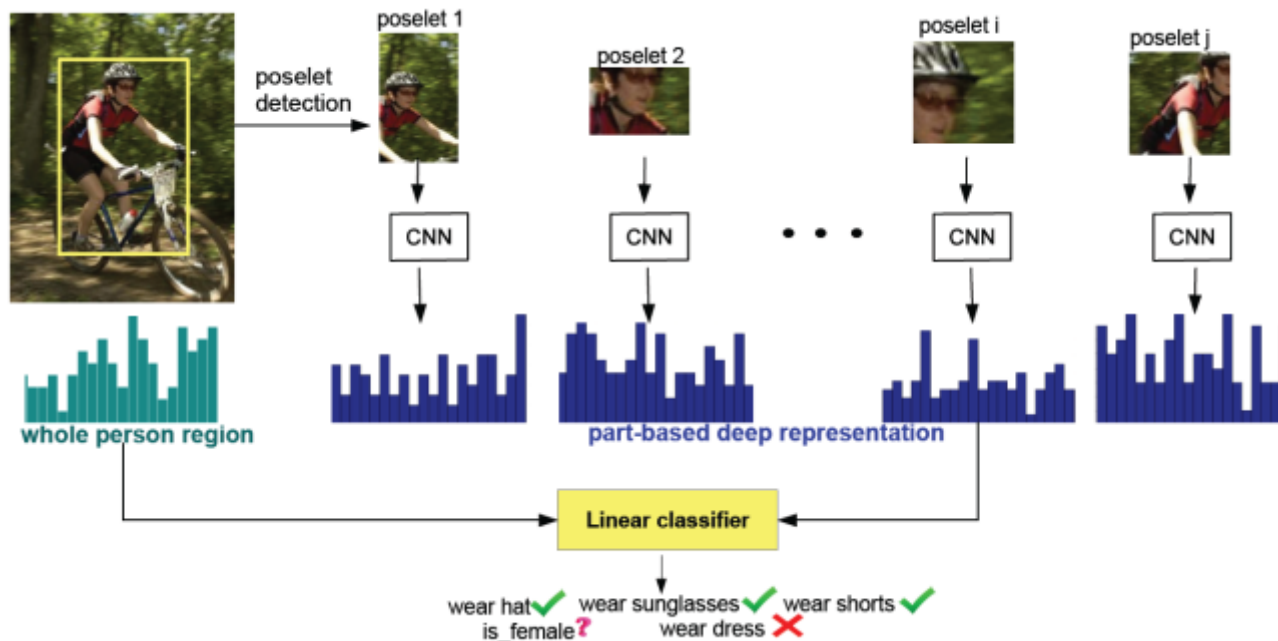
- Generalization over multiple tasks



Ali Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition, Arxiv 2014

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. ICLR 2014

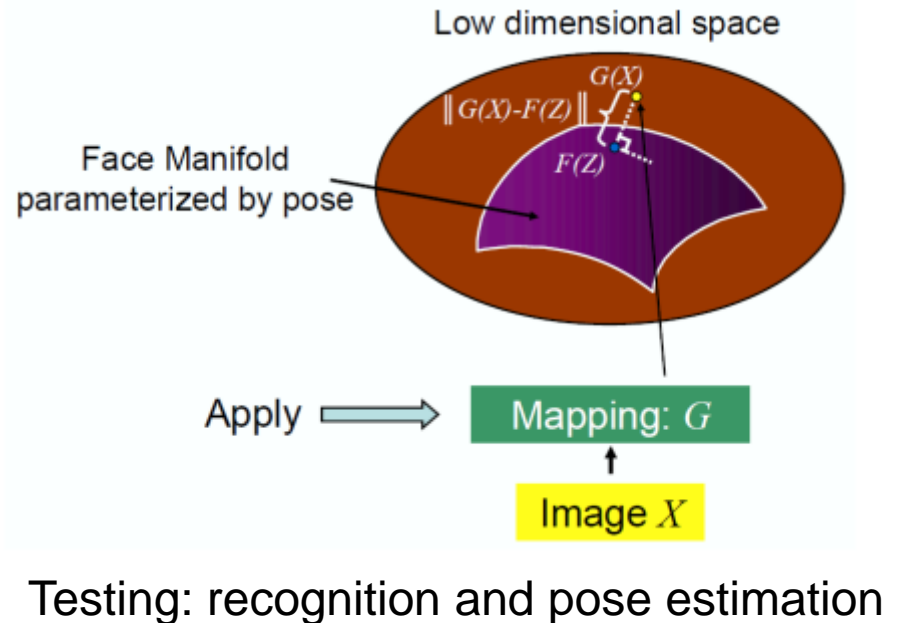
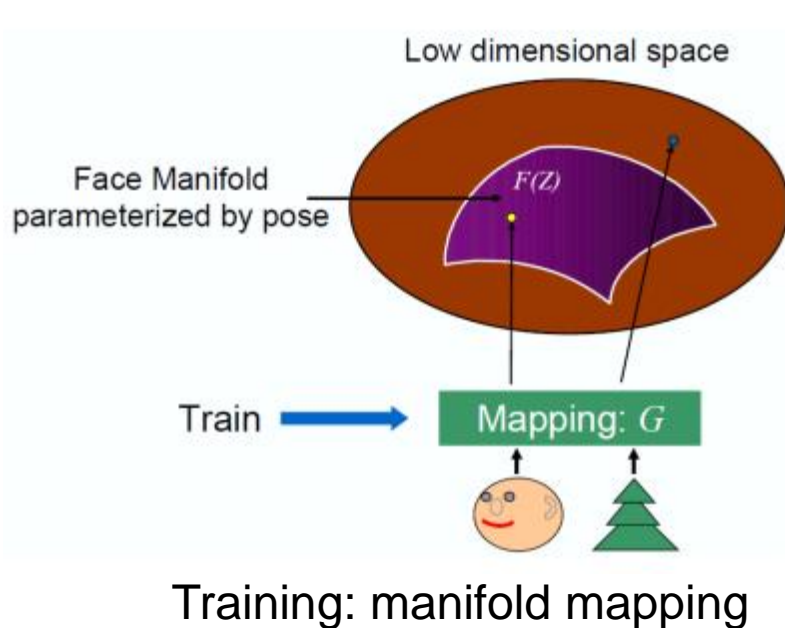
Multi-task deep learning for attribute prediction



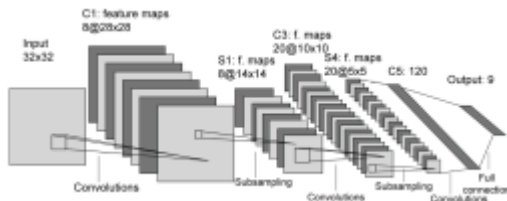
N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev,
PANDA: Pose Aligned Networks for Deep Attribute Modeling

Joint detection and pose estimation

- Main idea: project input data into low dimensional space (e.g., parameterized face-pose manifold)



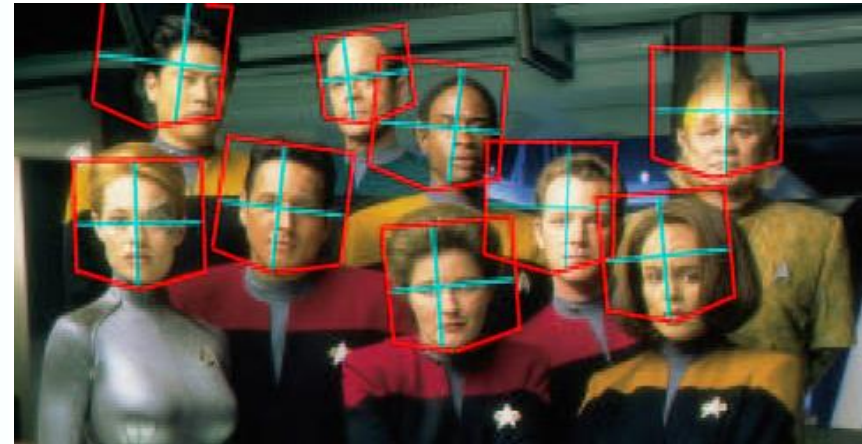
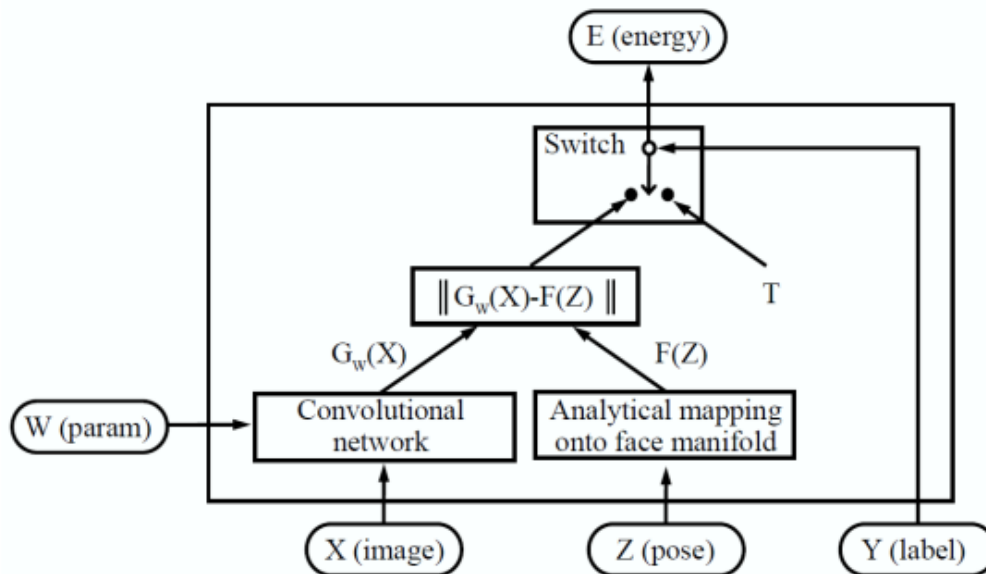
Mapping G :
Convnet



Oschady, LeCun, Miller, Synergistic Face Detection and Pose Estimation with Energy-Based Models, JMLR 2007

Joint detection and pose estimation

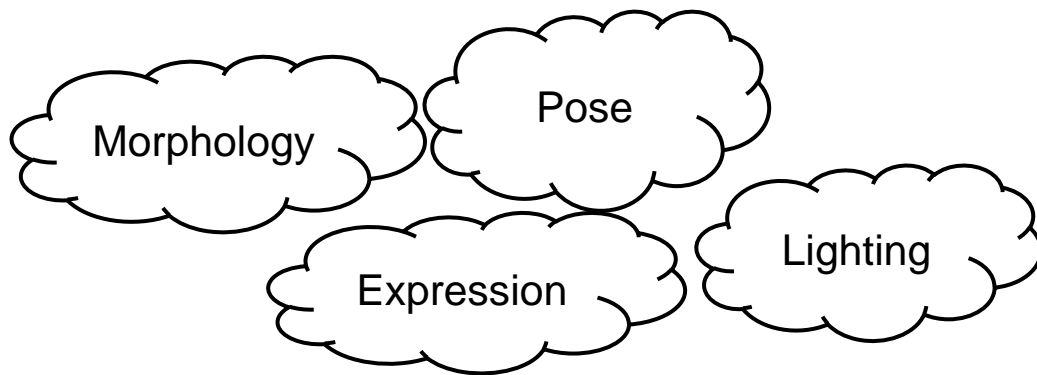
- Main idea: project input data into low dimensional space (e.g., parameterized face-pose manifold)



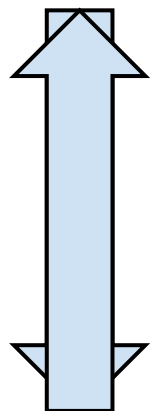
Disentangling Factors of Variation

Motivation

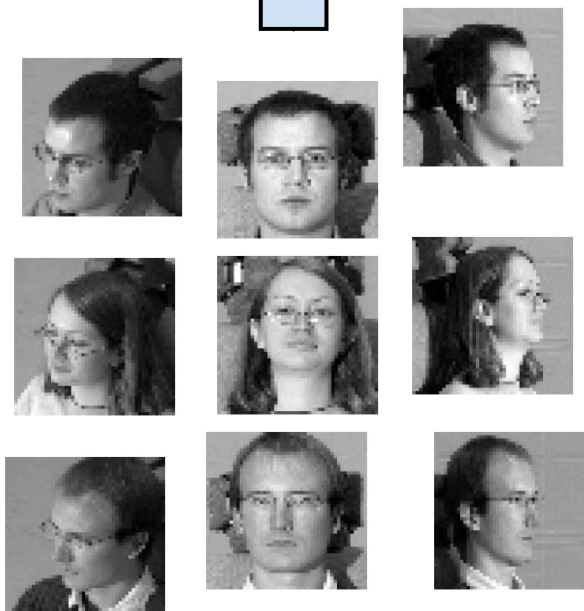
- A key challenge is to **tease apart the many factors of variation** that generate images by entangling
 - Morphology, pose, and expression for face images.
 - Pose, shape and illumination for 3D object images
- This challenge is closely related to **generalization ability** of learning algorithms
 - E.g., identifying faces under different facial expressions or viewpoints



Underlying Factors of Variation (unobserved)



How do we recover latent factors of variation?



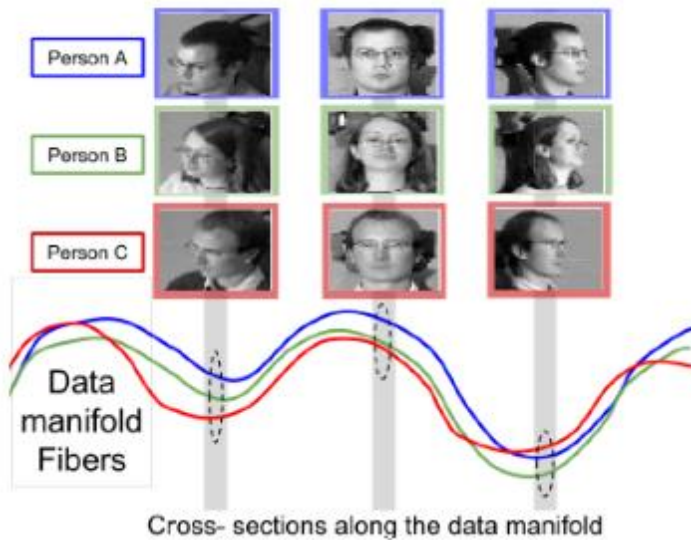
Complex sensory data

Related work:

Desjardins et al, Arxiv 2012;
Rifai et al., ECCV 2012;
Tennenbaum & Freeman, 2000;
Oschady, LeCun, Miller, 2007

Disentangling factors of variation

- Complex data can be generated by interaction between underlying factors of variation (manifolds)



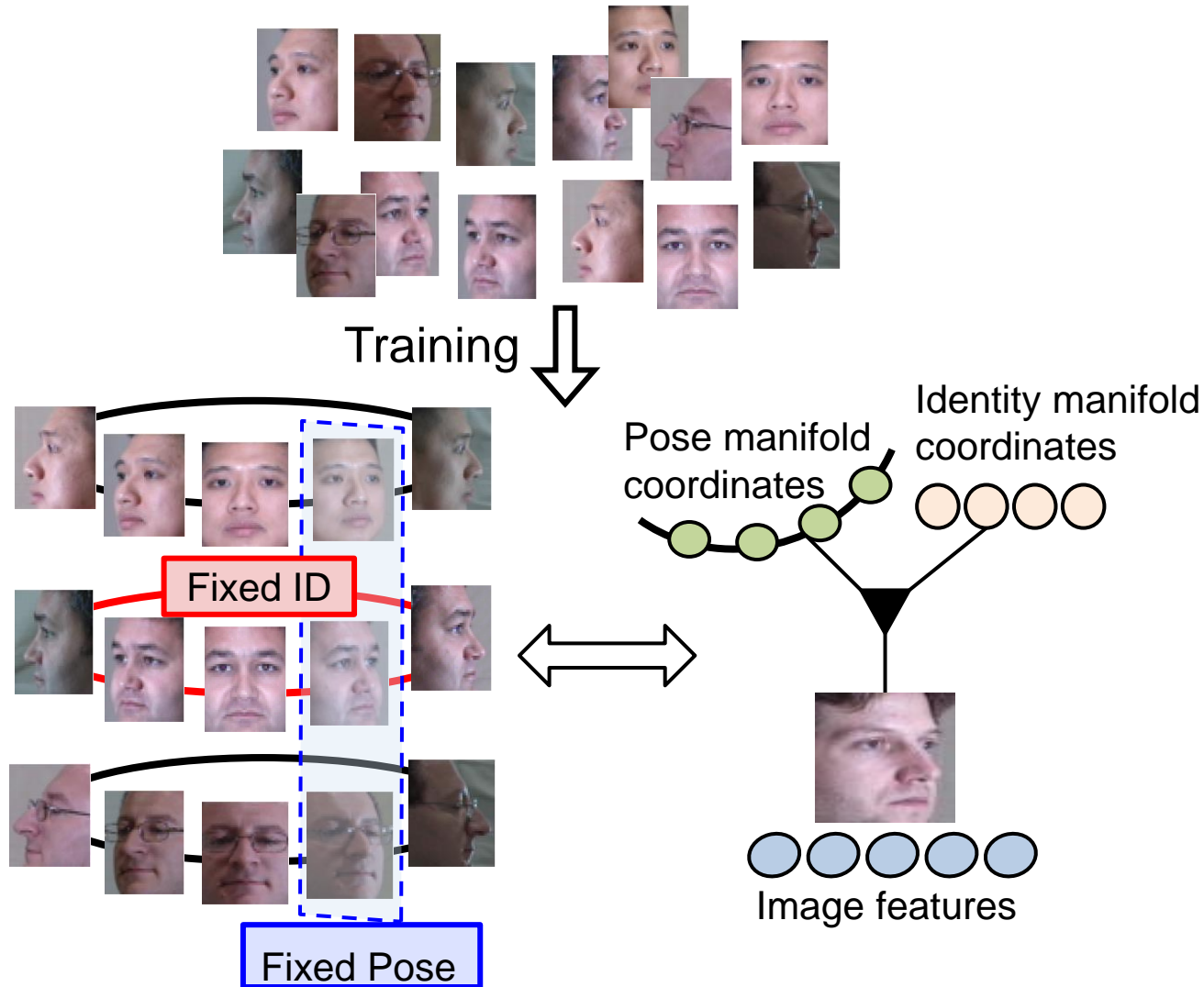
Abstract view of the data manifold as a collection of subject-specific *fibers* (or *low-dimensional manifold*).

- Moving along a fiber changes some factor of variation such as pose while preserving identity.
 - Hopping to another fiber at the same position changes identity while preserving pose.
- We would like to have separate coordinates for each factor of variation.**

Disentangling factors of variation



Disentangling factors of variation



Disentangling Boltzmann Machine

(Reed et al., ICML 2014)

Suppose we have observations \mathbf{v} ,
and two groups of hidden units \mathbf{h}
and \mathbf{m} :

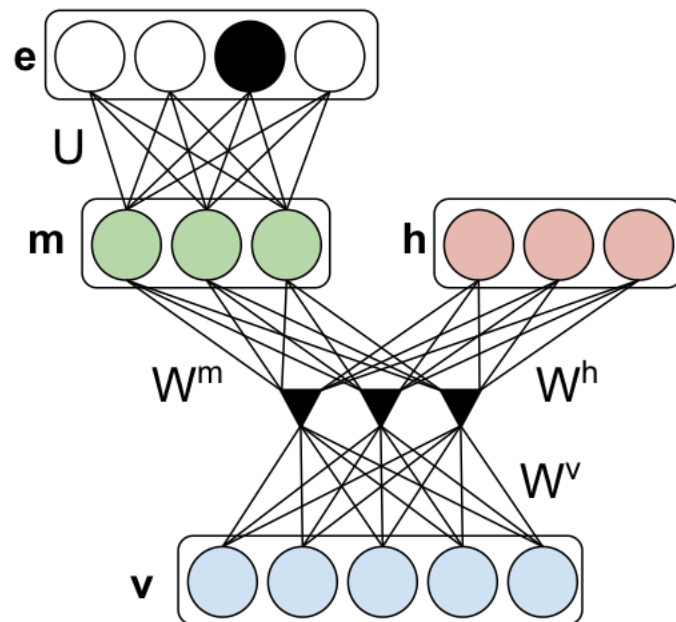
$$P(\mathbf{v}, \mathbf{m}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{m}, \mathbf{h}))$$

Ignoring the bias units, we can
write down the energy function as:

$$E(\mathbf{v}, \mathbf{m}, \mathbf{h}) = \underbrace{\sum_{ijk} W_{ijk} v_i h_j m_k}_{\text{3-way interaction}} + \sum_{ij} P_{ij}^h v_i h_j + \sum_{ik} P_{ik}^m v_i m_k$$

Low-rank
factorization

$$W_{ijk} = \sum_f W_{if}^{\mathbf{v}} W_{jf}^{\mathbf{m}} W_{kf}^{\mathbf{h}}$$



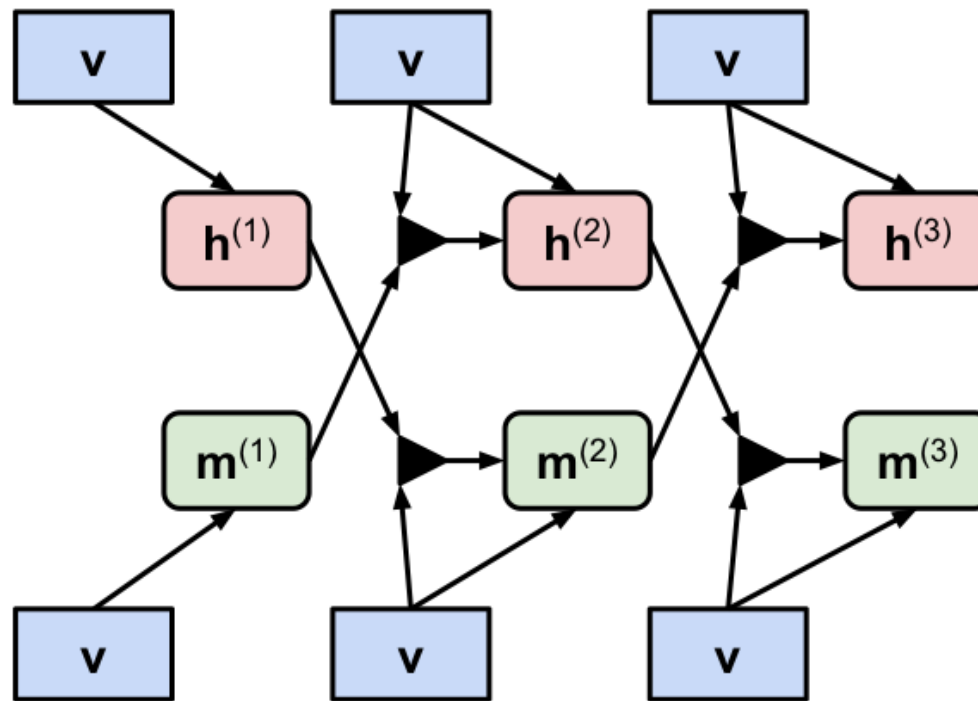
\mathbf{v} : input data

\mathbf{h} : identity hidden units

\mathbf{m} : emotion hidden units

\mathbf{e} : (optional) labels

Mean-field inference as RNN



$$\hat{h}_j = \text{sigmoid}\left(\sum_{ik} W_{ijk} v_i m_k + \sum_i P_{ij}^h v_i\right)$$

$$\hat{m}_k = \text{sigmoid}\left(\sum_{ij} W_{ijk} v_i h_j + \sum_i P_{ik}^m v_i\right)$$

Related work: Goodfellow et al., NIPS'13; Stoyanov, Ropson, Eisner, AISTATS'11

Weak supervision helps disentangling

Naive application of high-order BM doesn't work well.

Partial supervision

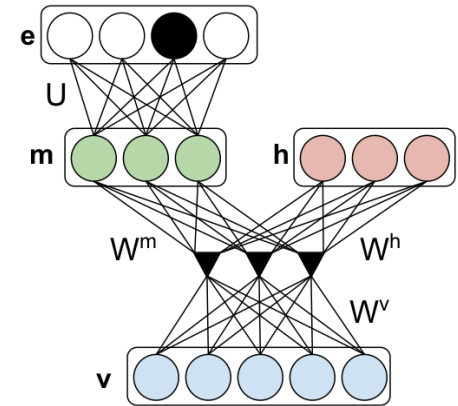
Partial labels (e.g. emotion label) can be used for constraining a specific factor during training.

Correspondence based learning

If we know two data points correspond (e.g. two images depict the same person), this can add useful regularizations in learning.

Exploiting correspondence

Example: Different pose, same identity.



Method 1: Clamping hidden units. This is accomplished using an augmented energy function that considers **pairs** of images $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)}$:

$$E_{\text{clamp}}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{m}^{(1)}, \mathbf{m}^{(2)}, \mathbf{h}) \\ = E(\mathbf{v}^{(1)}, \mathbf{m}^{(1)}, \mathbf{h}) + E(\mathbf{v}^{(2)}, \mathbf{m}^{(2)}, \mathbf{h})$$

Clamped hidden units
(e.g., identity units)

Exploiting correspondence

Example: Different identity, same expression.



Method: “Manifold” objective. For each pair, encourage corresponding features to be nearby, non-corresponding features to be apart.

$$\begin{aligned} \text{Want:} \quad & d(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}) \approx 0 \quad , \text{ if } (\mathbf{v}^{(1)}, \mathbf{v}^{(2)}) \in \mathcal{D}_{sim} \\ & d(\mathbf{h}^{(1)}, \mathbf{h}^{(3)}) \geq \beta \quad , \text{ if } (\mathbf{v}^{(1)}, \mathbf{v}^{(3)}) \in \mathcal{D}_{dis} \end{aligned}$$

$$\text{Objective:} \quad \lambda_1 d(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}) + \lambda_2 \max(0, \beta - d(\mathbf{h}^{(1)}, \mathbf{h}^{(3)}))$$

[Related work: Hadsell et al., 2006]

Disentangling Expression and ID

- Traversing along different “emotions”
 - Toronto face database (TFD)



Input

Disentangling Expression and ID

- Traversing along different “emotions”
 - Toronto face database (TFD)



Input



Disgust

Fear

Happy

Sad

Surprise

Neutral

Interpolation and Transfer: Multi-PIE

Interpolation



Generating neutral faces
from two different emotions

Emotion/Identity transfer



ID Expr. prediction

Interpolation and Transfer: Multi-PIE

Interpolation



Expr. 1 Prediction Expr. 2

Generating neutral faces
from two different emotions

Emotion/Identity transfer



ID Expr. prediction

Traversing along the viewpoint manifold

- Traversing along “viewpoints” (CMU-MultiPIE)

Input

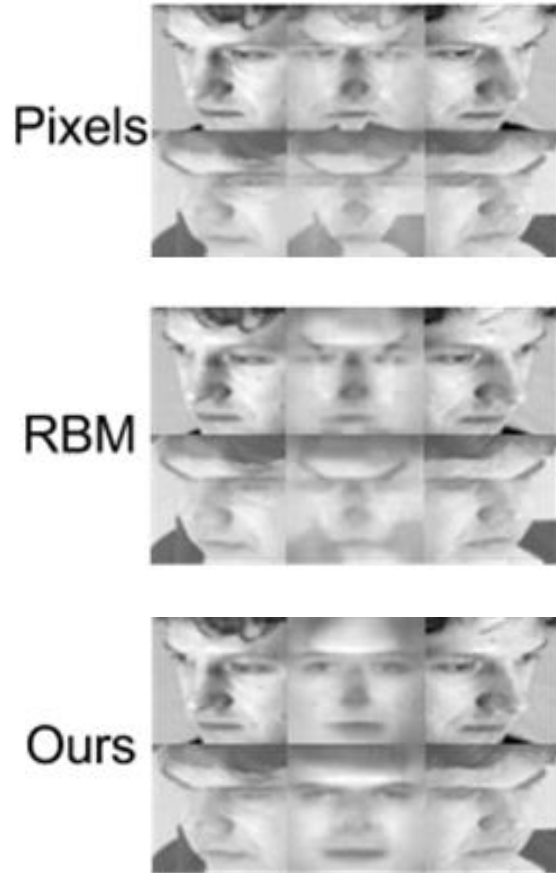


Generated (hallucinated) samples with different viewpoints



Interpolation and Transfer: Multi-PIE

Pose interpolation



Pose/Identity transfer



Discriminative Performance (TFD)

MODEL	EXPR. UNITS FOR EMOTION REC.	EXPR. UNITS FOR VERIFICATION	ID UNITS FOR EMOTION REC.	ID UNITS FOR VERIFICATION
NAIVE	79.50 \pm 2.17	0.835 \pm 0.018	79.81 \pm 1.94	0.878 \pm 0.012
LABELS (EXPR)	83.55 \pm 1.63	0.829 \pm 0.021	78.26 \pm 2.58	0.917 \pm 0.006
CLAMP (ID)	81.30 \pm 1.47	0.803 \pm 0.013	59.47 \pm 2.17	0.978 \pm 0.025
LABELS (EXPR) + CLAMP (ID)	82.97 \pm 1.85	0.799 \pm 0.013	59.55 \pm 3.04	0.978 \pm 0.024
MANIFOLD (BOTH)	85.43 \pm 2.54	0.513 \pm 0.011	43.27 \pm 7.45	0.951 \pm 0.025

- **Naive**: generative training only, with no correspondence or labels.
- **Labels (expr)**: generative training, also using 1-of-7 emotion labels when available. Unlabeled images are also used.
- **Clamp (ID)**: generative training with identity correspondence, but no label information.
- **Labels (expr) + Clamp (ID)**: both identity correspondence and emotion labels.
- **Manifold (both)**: both identity and emotion correspondence with manifold objective.

Correspondence-based learning significantly improves both discriminative performance and disentangling.

Discriminative Results: CMU Multi-PIE

TASK	POSE UNITS FOR POSE EST.	POSE UNITS FOR VERIFICATION	ID UNITS FOR POSE EST.	ID UNITS FOR VERIFICATION
NAIVE	96.60 \pm 0.23	0.583 \pm 0.004	95.79 \pm 0.37	0.640 \pm 0.005
LABELS (POSE)	98.07 \pm 0.12	0.485 \pm 0.005	86.55 \pm 0.23	0.656 \pm 0.004
CLAMP (ID)	97.18 \pm 0.15	0.509 \pm 0.005	57.37 \pm 0.45	0.922 \pm 0.003
LABELS (POSE) + CLAMP (ID)	97.68 \pm 0.17	0.504 \pm 0.006	49.08 \pm 0.50	0.934 \pm 0.002
MANIFOLD (BOTH)	98.20 \pm 0.12	0.469 \pm 0.005	8.68 \pm 0.38	0.975 \pm 0.002

- **Naive:** generative training only, with no correspondence or labels.
- **Labels (pose):** generative training, also using 1-of-15 pose labels when available.
- **Clamp (ID):** generative training with identity correspondence, but no label information.
- **Labels (pose) + Clamp (ID):** both identity correspondence and pose labels.
- **Manifold (both):** both identity and pose correspondence with manifold objective.

Correspondence-based learning significantly improves both discriminative performance and disentangling.

Summary

- Deep learning features achieve significantly improved generalization performance in multi-task learning and transfer learning tasks
- Disentangling factors of variation holds promise in allowing deep learning algorithm jointly infer complex variabilities