Bayesian multimodeling: knowledge transfer

MIPT

2023

Related tasks

- Knowledge distillation
- Multidomain adaptation/learning
- Transfer learning
- Multiview learning
- Mulitmodal learning

Knowledge distillation [Hinton et al., 2015]

Given a teacher model. Knowledge distillation is a transfer of the knowledge from the teacher model to a small model that is more suitable for deployment.

Teacher model

Information

$\rightarrow$

Student model

Distillation and privileged information

General approach for distillation [Lopez-Paz et al., 2015]:

- Train teacher using extendnd dataset X^*, y^*;
- train student usign teacher output and basic dataset X, y.

If $X = X^*$, we get a distillation.
Generally X^* can contain not only object features, but other information.
Example: object similarity.

Hinton distillation

$$\lambda \log p\left(\mathrm{y}|\mathbf{SM}\left(\mathrm{f}_{\mathrm{student}}(\mathrm{X}) / T\right) + (1-\lambda) \log p\left(\mathbf{SM}\left(\mathrm{f}_{\mathrm{teacher}} / T\right)|\mathbf{SM}\left(\mathrm{f}_{\mathrm{student}}(\mathrm{X}) / T\right),\right.\right.$$

where $\mathrm{f}_{\mathrm{student}}$ and $\mathrm{f}_{\mathrm{teacher}}$ are student and teacher outputs, T is a temperature:

- When $T \to \infty$, we get uniform distribution;
- With T is high enough and outputs are zero-mean we get l_2 minimization of distance between student and teacher logits.
- When $T \to 0$, we get one-hot labels from teacher.

Homogeneous model distillation

Teacher layer 1

$\rightarrow$

Student layer 1

Teacher layer 2

$\rightarrow$

Student layer 2

Teacher layer 3

$\rightarrow$

Student layer 3

TinyBERT: example

$$\mathcal{L}_{\text{attn}} = \frac{1}{h}\sum_{i=1}^{h} \texttt{MSE}(\boldsymbol{A}_i^S, \boldsymbol{A}_i^T),$$

$$\mathcal{L}_{\text{hidn}} = \texttt{MSE}(\boldsymbol{H}^S\boldsymbol{W}_h, \boldsymbol{H}^T),$$

$$\mathcal{L}_{\text{embd}} = \texttt{MSE}(\boldsymbol{E}^S\boldsymbol{W}_e, \boldsymbol{E}^T),$$

$$\mathcal{L}_{\text{pred}} = \texttt{CE}(\boldsymbol{z}^T/t, \boldsymbol{z}^S/t),$$

Heterogenous model distillation

- There can be different number of layers
- Layers can have different functionality

Transformer

???

LSTM

TextBrewer: naive distillation example

Model	MNLI		SQuAD		CoNLL-2003
	m	mm	EM	F1	F1
$\text{BERT}_{\text{BASE}}$	83.7	84.0	81.5	88.6	91.1
Public					
DistilBERT	81.6	81.1	79.1	86.9	-
TinyBERT	80.5	81.0	-	-	-
+DA	82.8	82.9	72.7	82.1	-
TextBrewer					
BiGRU	-	-	-	-	85.3
T6	83.6	84.0	80.8	88.1	90.7
T3	81.6	82.5	76.3	84.8	87.5
T3-small	81.3	81.7	72.3	81.4	78.6
T4-tiny	82.0	82.6	73.7	82.5	77.5
+DA	-	-	75.2	84.0	89.1

Model	# Layers	Hidden size	Feed-forward size	# Parameters	Relative size
$\text{BERT}_{\text{BASE}}$ (teacher)	12	768	3072	108M	100%
T6	6	768	3072	65M	60%
T3	3	768	3072	44M	41%
T3-small	3	384	1536	17M	16%
T4-tiny	4	312	1200	14M	13%
BiGRU	1	768	-	31M	29%

Cosine distribution: naive distillation example

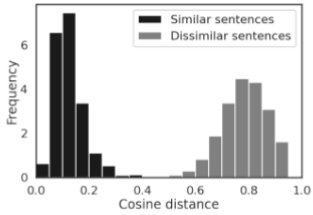


Figure 2a. The distribution of distances between pairs of similar and dissimilar sentences with different phrase embedding models: LaBSE model.

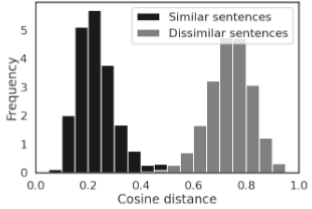


Figure 2b. The distribution of distances between pairs of similar and dissimilar sentences with different phrase embedding models: LSTM model.

Probabilistic knowledge transfer [Passalis et al., 2018]

Main idea: Given two homogenuos models with one hidden layer $\mathrm{h}_t, \mathrm{h}_s$.
Distillation problem:

$$\mathrm{KL}\left(p\left(\mathrm{H}_s, \mathrm{H}_t\right)\right) \rightarrow \min .$$

Alternative problem statement:

$$I\left(\mathrm{H}_s, \mathrm{y}\right) = I\left(\mathrm{H}_t, \mathrm{y}\right),$$

where $I(a, b) = \mathrm{KL}\left(p(a, b) | p(a)p(b)\right)$.
Approximation:

$$IQ = \mathrm{E}_{\mathrm{H},\mathrm{y}} || p(\mathrm{H}_t, \mathrm{y}) - p(\mathrm{H}_t)p(\mathrm{y}) ||_2^2.$$

PKT: probability estimation

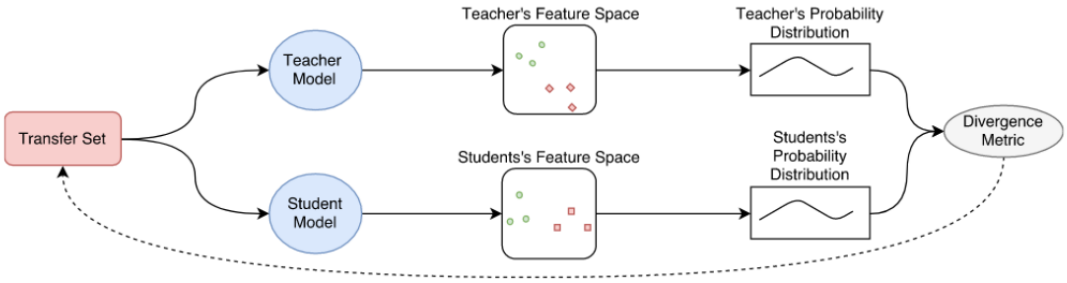
$$IQ(\mathrm{h}_t, \mathrm{y}) \approx \sum_y \sum_{\mathrm{h}_t} p(\mathrm{h}_t, y)^2 + \left(p(\mathrm{h}_t)p(y)\right)^2 - 2p(\mathrm{h}_t)p(y)p(\mathrm{h}, y).$$

$p(y)$ is esitmated empirically, $p(\mathrm{h})$ is estimated using KDE:

$$p(\mathrm{h}) \propto \sum_{\mathrm{h}'} K(\mathrm{h}, \mathrm{h}'),$$

$p(\mathrm{h}, y)$ is estimated in the same way, but the sum is only over objects of the class y.

PKT


Transfer Set
Teacher Model
Student Model
Teacher's Feature Space
Students's Feature Space
Teacher's Probability Distribution
Students's Probability Distribution
Divergence Metric

Can we extend this approach for heterogenuos models?

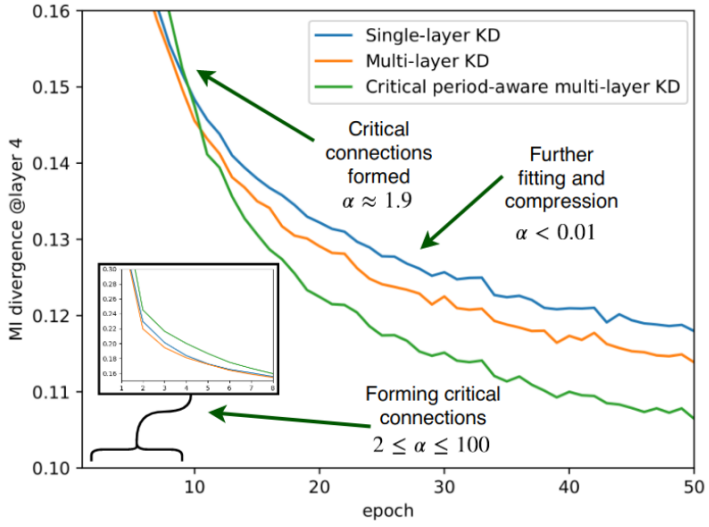
- Create a «heavy» model that
  - homogenous to student
  - has larger capacity
- Distil from teacher to heavy model (without one-to-one layer alignment)
- Distil from heavy model to student model

Results

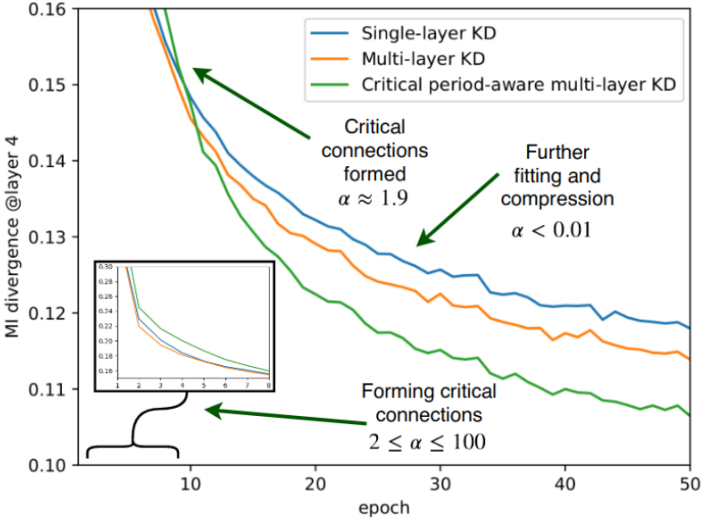
Table 1. Metric Learning Evaluation: CIFAR-10

Method	mAP (e)	mAP (c)	top-100 (e)	top-100 (c)
Baseline Models				
Teacher (ResNet-18)	87.18	90.47	92.15	92.26
Aux. (CNN1-A)	62.12	66.78	73.72	75.91
With Constrastive Supervision				
Student (CNN1)	47.69	48.72	57.46	58.50
Hint.	43.56	48.73	**60.44**	**62.43**
MKT	45.34	46.84	55.89	57.10
PKT	48.87	49.95	58.44	59.48
Hint-H	43.24	47.46	58.97	61.07
MKT-H	44.83	47.12	56.28	57.90
PKT-H	48.69	50.09	58.71	60.20
Proposed	**49.55**	**50.82**	59.50	60.79
Without Constrastive Supervision				
Student (CNN1)	35.30	39.00	55.87	58.77
Distill.	37.39	40.53	56.17	58.56
Hint.	43.99	48.99	60.69	62.42
MKT	36.26	38.20	50.55	52.72
PKT	48.07	51.56	60.02	62.50
Hint-H	42.65	46.46	58.51	60.59
MKT-H	41.16	43.99	55.10	57.63
PKT-H	48.05	51.73	60.39	63.01
Proposed	**49.20**	**53.06**	**61.54**	**64.24**

Intermediate layer importance



Variational Information Distillation for Knowledge Transfer (Ahn et al., 2019)



Bayesian deep learning model distillation (Grabovoy, Strijov, 2021)

- Teacher is trained using ELBO
- Student is trained using prior equal to student posterior
- If teacher and student are heterogenous we additionally solve a problem of distribution alignment pruning irrelevant parameters

Distillation without dataset (Lopes et al., 2017)

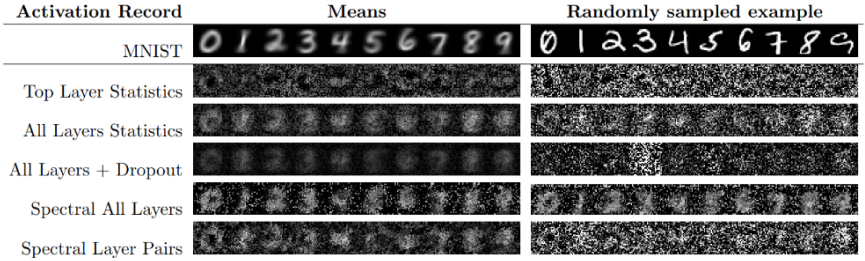
Challenge: for distillation we need to store dataset.
What if our dataset contains terrabytes of data?
Solution: we will try to restore objects from teacher model.

Distillation without dataset (Lopes et al., 2017)

Simple way:

- Store teacher output distribution for each layer (mean and covariance)
- For distilation step: $b \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$;
- For student model f : $x_0 = \arg\min_x ||f(x) - b||_2^2$;
- Distil using x_0.

Distillation without dataset (Lopes et al., 2017)



Distillation without dataset (Lopes et al., 2017)

Table 5: Accuracies of the AlexNet model and CelebA dataset for each procedure.

Model Name	**Procedure**	**Accuracy on test set**
AlexNet	Train on CelebA	80.82%
AlexNet-half	Train on CelebA	81.59%
AlexNet-half	Knowledge Distillation [8]	69.53%
AlexNet-half	Top Layer Statistics	54.12%
	All-Layers Spectral	77.56%
	Layer-Pairs Spectral	76.94%

Dataset shift

Dataset shift is an event when distribuition $p(\mathrm{X}, \mathrm{y})$ significantly differ for the training and test/inference phases.

- Covariate shift — difference in $p(\mathrm{X})$
- Prior probability shift — difference in $p(\mathrm{y})$
- Concept shift — difference in $p(\mathrm{y}|\mathrm{X})$

Dataset shift

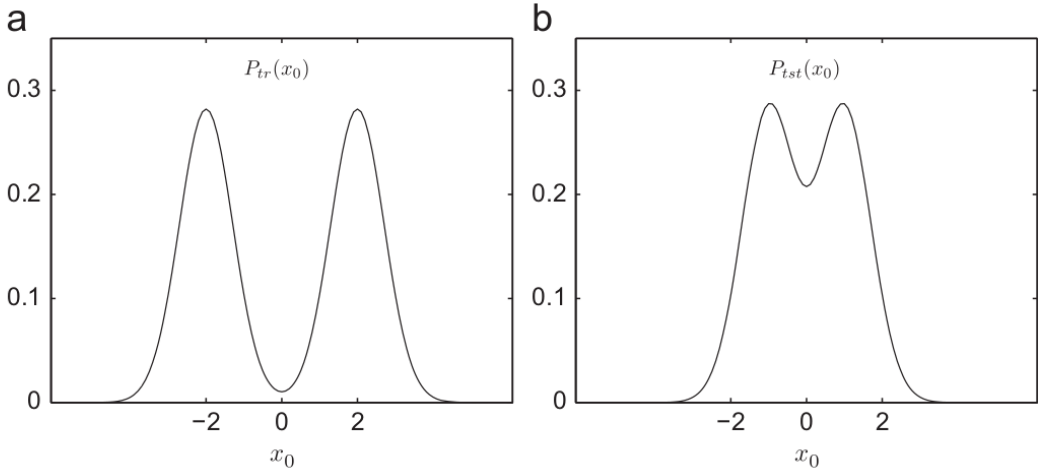
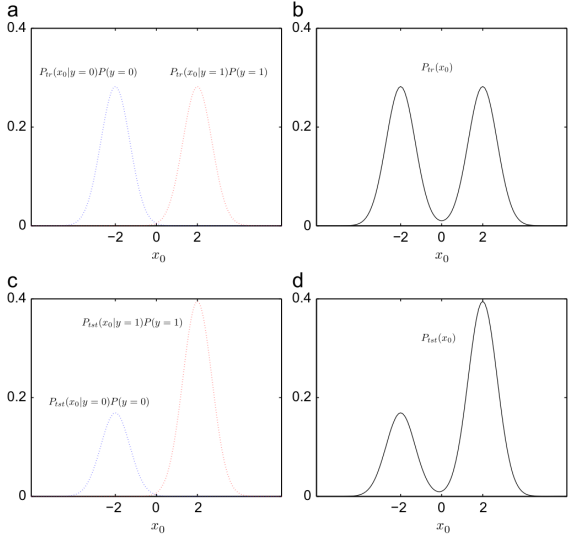


Fig. 1. Covariate shift: $P_{tst}(y|x_0) = P_{tr}(y|x_0)$ and $P_{tr}(x_0) \neq P_{tst}(x_0)$. (a) Training data and (b) test data.

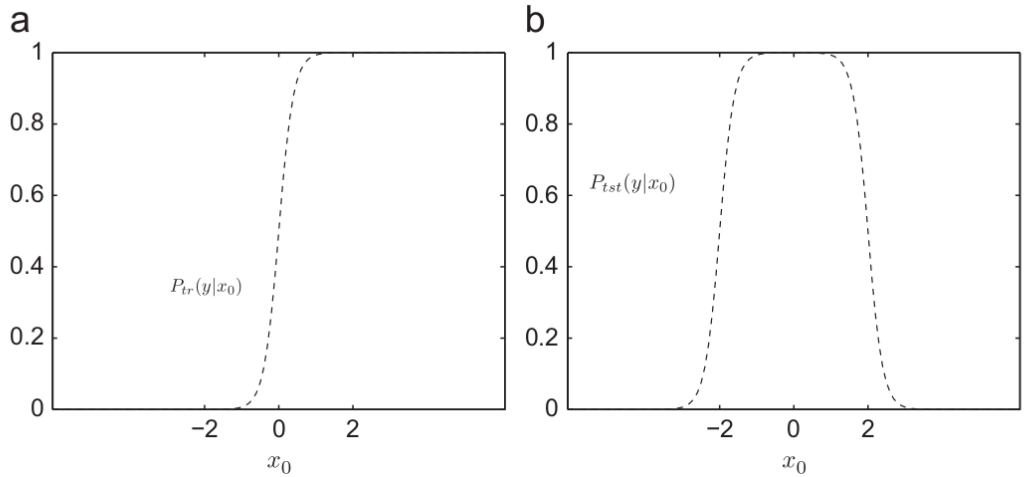
Moreno-Torres et al., 2012

Dataset shift



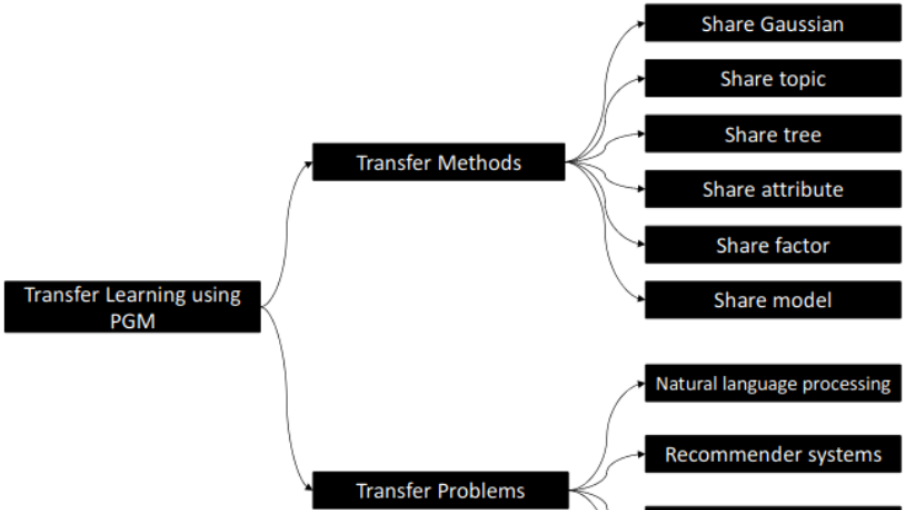
Moreno-Torres et al., 2012

Dataset shift

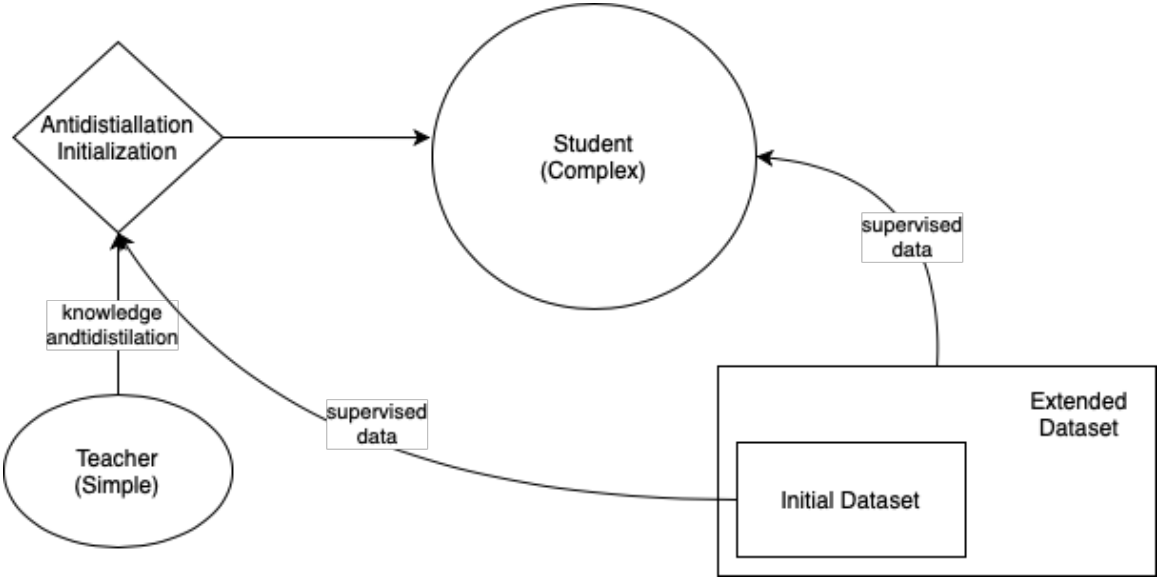


Moreno-Torres et al., 2012

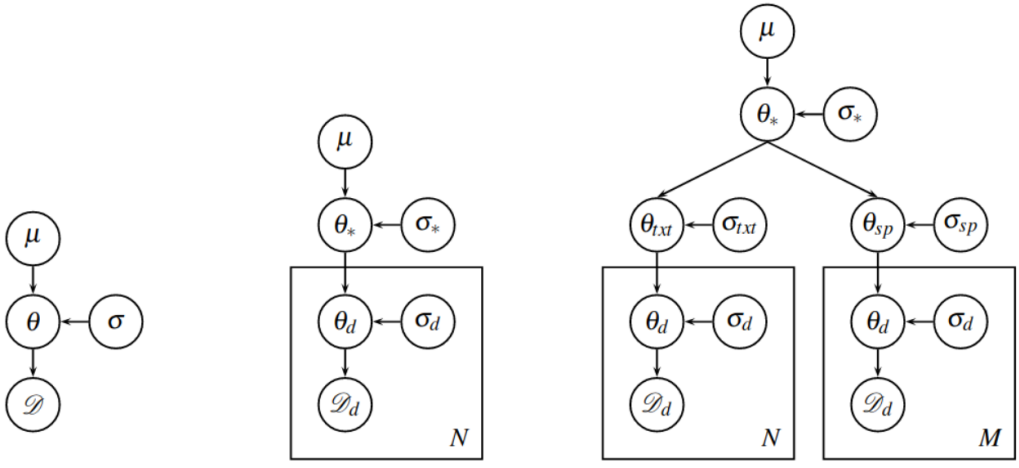
Knwoledge transfer: method scheme



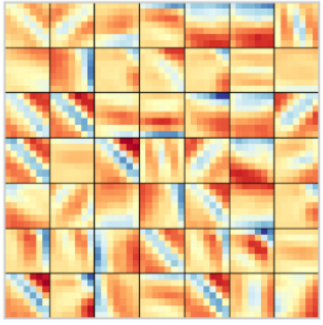
Antidistillation



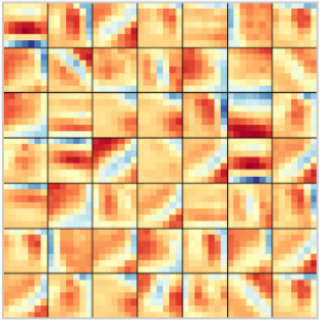
Share prior: Hierarchical Bayesian Domain Adaptation



The deep weight prior: Atanov et al., 2019



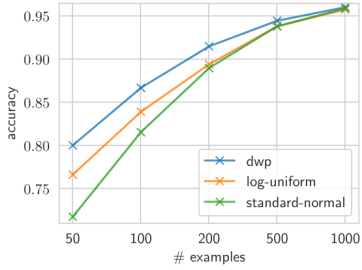
(b) Learned filters



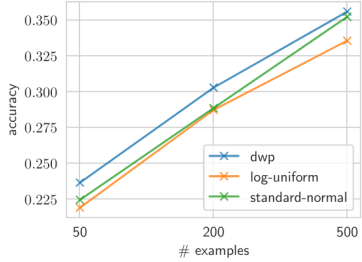
(c) Samples from DWP

The distribution can be modeled by complex models and can generate rather informative samples!

The deep weight prior: Atanov et al., 2019

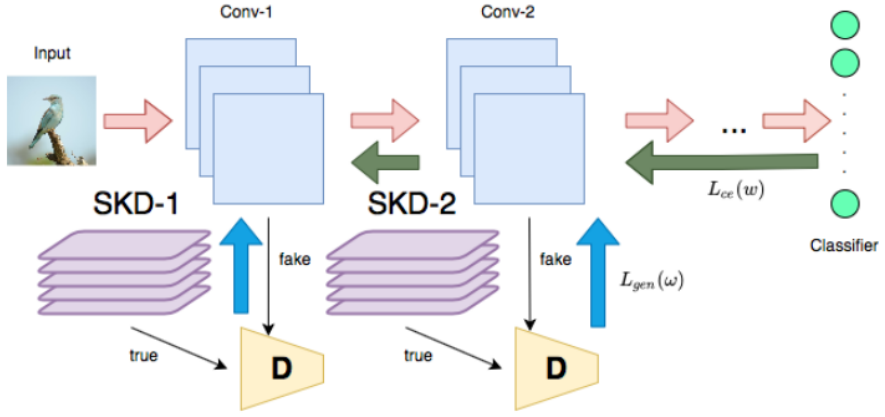


(a) Results for MNIST

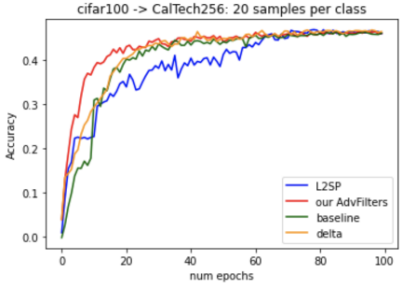


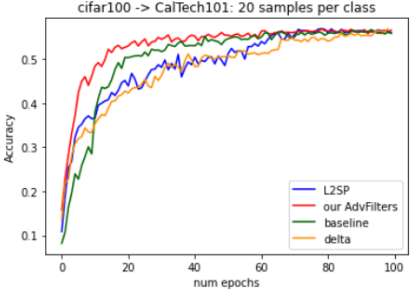
(b) Results for CIFAR-10

Deep weight prior for distillation, Kolesov 2022

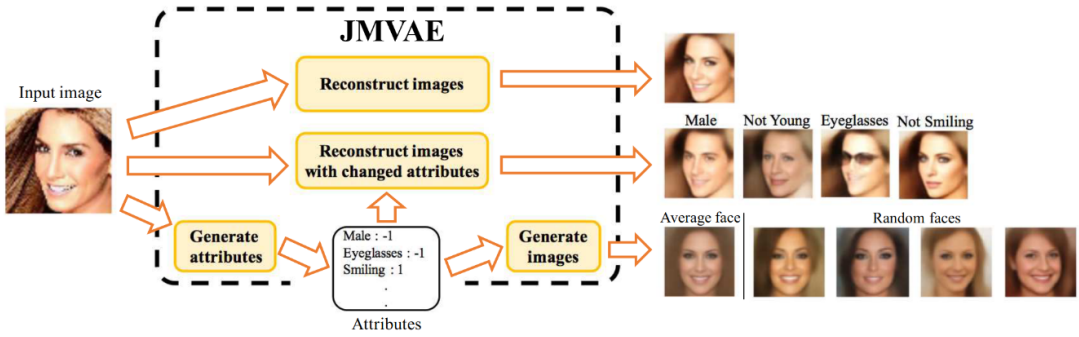


Deep weight prior for distillation, Kolesov 2022



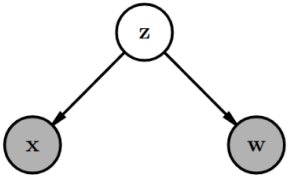


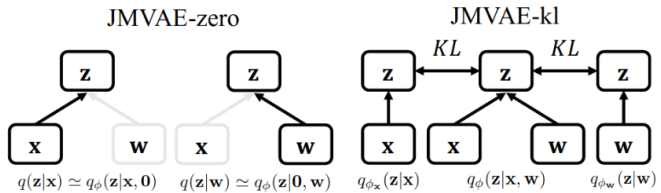
Share topic: Joint multimodal learning with deep generative models



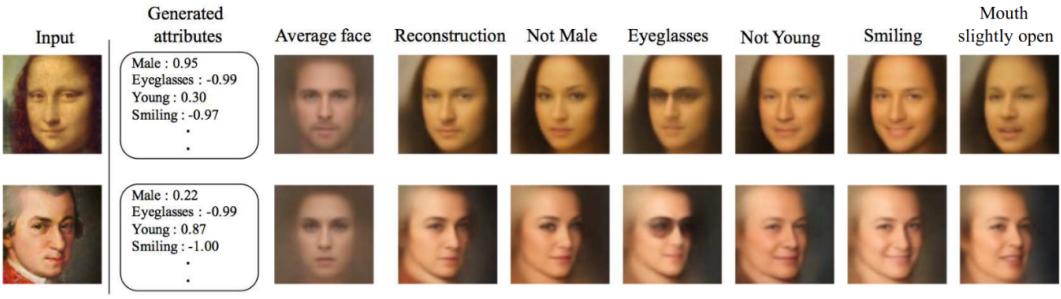
Share topic: Joint multimodal learning with deep generative models

$$
\begin{aligned}
\mathcal{L}_{JM}(\mathbf{x}, \mathbf{w}) &= E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{p_\theta(\mathbf{x}, \mathbf{w}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})}] \\
&= -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})||p(\mathbf{z})) \\
&\quad + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_{\theta_\mathbf{w}}(\mathbf{w}|\mathbf{z})].
\end{aligned}
$$

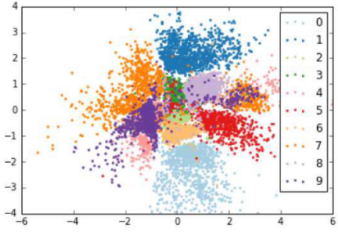




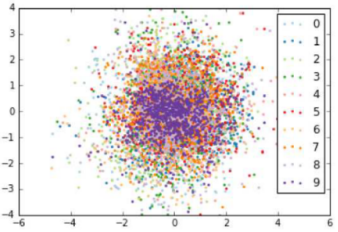
Share topic: Joint multimodal learning with deep generative models



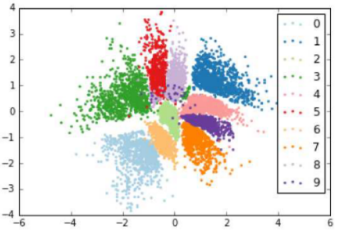
Share topic: Joint multimodal learning with deep generative models



(a) VAE

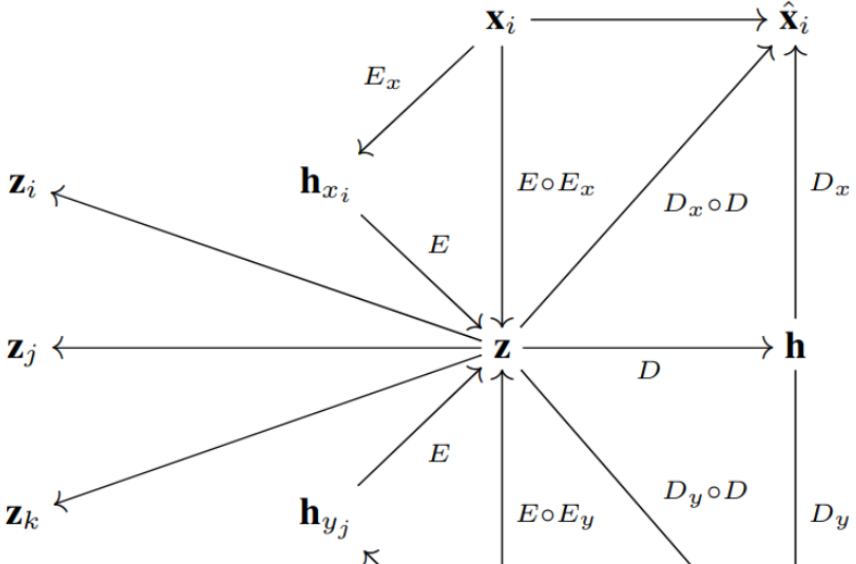


(b) CVAE

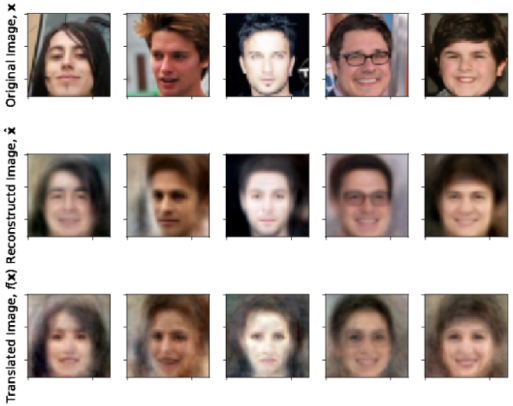


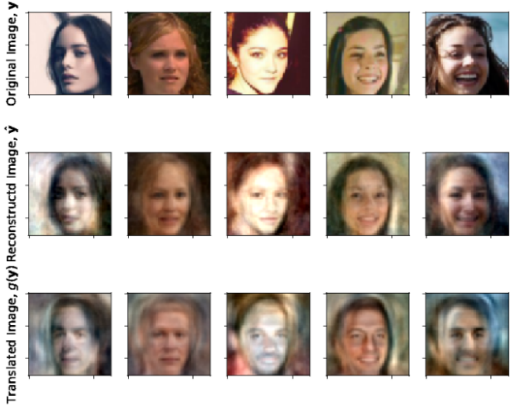
(c) JMVAE

Share topic:Variational learning across domains with triplet information

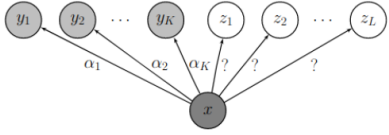


Share topic:Variational learning across domains with triplet information

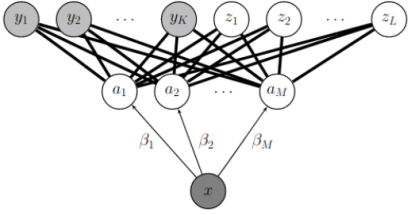




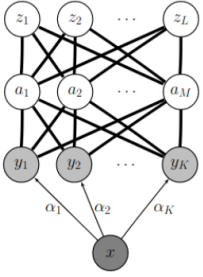
Share attributes



(a) Flat multi-class classification



(b) Direct attribute prediction (DAP)



(c) Indirect attribute prediction (IAP)

Share attributes

otter

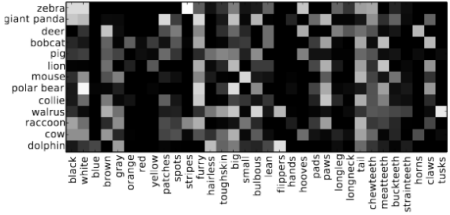
black:	yes
white:	no
brown:	yes
stripes:	no
water:	yes
eats fish:	yes

polar bear

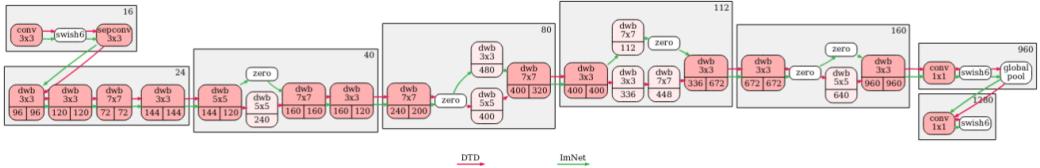
black:	no
white:	yes
brown:	no
stripes:	no
water:	yes
eats fish:	yes

zebra

black:	yes
white:	yes
brown:	no
stripes:	yes
water:	no
eats fish:	no



Share models



References

- Hinton G. et al. Distilling the knowledge in a neural network //arXiv preprint arXiv:1503.02531. – 2015. – T. 2. – No. 7.
- Lopez-Paz D. et al. Unifying distillation and privileged information //arXiv preprint arXiv:1511.03643. – 2015.
- Jiao X. et al. Tinybert: Distilling bert for natural language understanding //arXiv preprint arXiv:1909.10351. – 2019.
- Yang Z. et al. Textbrewer: An open-source knowledge distillation toolkit for natural language processing //arXiv preprint arXiv:2002.12620. – 2020.
- Bakhteev Oleg et al. Cross-language plagiarism detection: a case study of European universities academic works // ENAI. - 2021.
- Passalis N., Tefas A. Learning deep representations with probabilistic knowledge transfer //Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – С. 268-284.
- Ahn S. et al. Variational information distillation for knowledge transfer //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2019. – С. 9163-9171.
- Грабовой А. В., Стрижов В. В. Байесовская дистилляция моделей глубокого обучения //Автоматика и телемеханика. – 2021. – No. 11. – С. 16-29.

References

- Янина А. О., Воронцов К. В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге //Машинное обучение и анализ данных. – 2016. – Т. 2. – No. 2. – С. 173-186.
- Кузнецова М.В. Вариационное моделирование правдоподобия с триплетными ограничениями в задачах информационного поиска. 2021 (диссертация)
- Moreno-Torres J. G. et al. A unifying view on dataset shift in classification //Pattern recognition. – 2012. – Т. 45. – No. 1. – С. 521-530.
- Xuan J., Lu J., Zhang G. Bayesian Transfer Learning: An Overview of Probabilistic Graphical Models for Transfer Learning //arXiv preprint arXiv:2109.13233. – 2021.
- Finkel J. R., Manning C. D. Hierarchical bayesian domain adaptation //Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics. – 2009. – С. 602-610.
- Sultan M. A., Boyd-Graber J., Sumner T. Bayesian supervised domain adaptation for short text similarity //Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2016. – С. 927-936.
- Suzuki M., Nakayama K., Matsuo Y. Joint multimodal learning with deep generative models //arXiv preprint arXiv:1611.01891. – 2016.
- Kuznetsova R., Bakhteev O., Ogaltsov A. Variational learning across domains with triplet information //arXiv preprint arXiv:1806.08672. – 2018.
- Lampert C. H., Nickisch H., Harmeling S. Learning to detect unseen object classes by between-class attribute transfer //2009 IEEE conference on computer vision and pattern recognition. – IEEE, 2009. – С. 951-958.
- Wang Q. et al. Multi-path neural networks for on-device multi-domain visual classification //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. – 2021. – С. 3019-3028
- Atanov, Andrei, et al. "The deep weight prior."arXiv preprint arXiv:1810.06943 (2018).
- Kolesov A. An adversarial method for neural network fine-tuning for transfer learning problem, Master thesis, 2022.