EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-Naturwissenschaftliche Fakultät

Lernbasierte Computer Vision

Bachelorarbeit

A Comparison of Synthetic-to-Real Domain Adaptation Techniques

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik
Lernbasierte Computer Vision
Peter Trost, peter.trost@student.uni-tuebingen.de, 2019

Bearbeitungszeitraum: 24.05.2019-23.09.2019

Betreuer/Gutachter: Prof. Dr. Andreas Geiger, Universität Tübingen

Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Peter Trost (Matrikelnummer 4039682), June 18, 2019

Abstract

Template

Acknowledgments

If you have someone to Acknowledge ;)

Contents

1. Introduction

What is this all about?

Cite like this: [GPAM+14]

1.1. Problem Statement

TODO: what you have to do here :)

2. Related Work

2.1. Adapting Visual Category Models to New Domains

see [SKFD10]

Notes while reading:

2.1.1. Abstract

- one of the first studies of domain shift in context of object recognition
- method that adapts object models acquired in a particular visual domain to new imaging conditions by learning a transformation that minimizes the effect of domain-induced changes in the feature distribution **TODO: is copy paste, rephrase this!**
- supervised learning
- no labeled examples in the new domain needed
- could also be applied to non-image data
- authors also contribute a freely available multi-domain object database

2.1.2. Introduction

- kernel-based, nearest-neighbor classifiers (**TODO: look these up**) often fail on other visual domains than the one trained on
- often want to label *target* visual domain that doesn't have labels yet while having access to *source* domain that has labeled examples
- insufficient using object classifiers trained on source domain **TODO: include Figure 1, look up SVM[-bow] and NBNN**
- domain shift can affect feature distribution and cause the classifier to fail its prediction
- causes of visual domain shift include changes in camera, image resolution, lighting, background, viewpoint, post-processing **TODO: rephrase!**

- introduce domain adaptation technique based on cross-domain transformations
- key idea: regularized non-linear transformation that maps points in the source domain (green) closer to those in the target domain (blue) using supervised data from both domains. The input consists of labeled pairs of inter-domain examples that are known to be either similar (black lines) or dissimilar (red lines). The output is the learned transformation, which can be applied to previously unseen test data points. **TODO: include Figure 2, rephrase!**
- **TODO: look up '[theoretic] metric learning'**

2.1.3. Domain Adaptation Using Regularized Cross-Domain Transforms

general domain adaptation model in linear setting:
let source domain be $\mathcal{A}$ and target domain $\mathcal{B}$. Vectors $\mathbf{x} \in \mathcal{A}$, $\mathbf{y} \in \mathcal{B}$. Learn transformation W from $\mathcal{B}$ to $\mathcal{A}$ (and W^T from $\mathcal{A}$ to $\mathcal{B}$). Let dimensionality of $\mathbf{x}$ be d_A and of $\mathbf{y}$ be d_B then the transformation matrix W is $d_A \times d_B$. Resulting inner product similarity function between $\mathbf{x}$ and the transformed $\mathbf{y}$ as

$$\text{sim}_W(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T W \mathbf{y}$$

to avoid overfitting use regularization function for W denoted as $r(W)$.

TODO: lookup Mahalanobis metric learning method, information theoretic metric learning (ITML)

($||W||$: square root of sum of squares of elements)

2.2. Adversarial Discriminative Domain Adaptation

see [THSD17]

TODO: Look up cross entropy loss

2.3. Adversarial Dropout Regularization

see [SUHS17]

experiments on p4d dataset and VisDA-classification
TODO: read this again for the benchmarks and comparison of domain adaptation techniques

2.4. Adversarial Feature Learning

see [DKD16]

TODO: look up jensen-shannon divergence

2.5. Adversarially Learned Inference

see [DBP+16]

3. Datasets

4. Domain Adaptation Techniques

4.1. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

see [ZPIE17]

- image-to-image translation: extracting characteristics of an image and translating it to another style while preserving the characteristics (rgb to greyscale, painting to photo,..)
- special in this approach: no paired images necessary (datasets with paired images are far more expensive)
- create mapping $G : X \to Y$ from source domain X to target domain Y
- The Generator has to trick the discriminator into believing $G(x), x \in X$ is actually a real sample y from the target domain Y (matches distribution $p_{\text{data}}(y)$)
- problem of mode collapse: any inputimage will be translated to the same output image
- add cycle-consistency constraint: create mapping $F : Y \to X$ and add contraint $F(G(x)) \overset{!}{\approx} x$
- objective for mapping/generator G and discriminator D_Y:
  $\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[1 - \log D_Y(G(x))]$
- analogous for mapping/generator F and discriminator D_X
- generators try to minimize the objective, discriminators try to maximize it
- cycle consistency loss:
  $\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$
- full objective:
  $\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F)$
- solve: $G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y)$

4.1.1. Training Details

- for $\mathcal{L}_{\text{GAN}}$ replaced negative log-likelihood objective by least-squares loss → more stable during training and generates higher quality results
- for GAN loss $\mathcal{L}_{\text{GAN}}(G, D, X, Y)$ they train
  G to minimize $\mathbb{E}_{x \sim p_{\text{data}}(x)}[(D(G(x)) - 1)^2]$
  and D to minimize $\mathbb{E}_{y \sim p_{\text{data}}(y)}[(D(y) - 1)^2] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[D(G(x))^2]$
- reduce model oscillation **TODO: add Shrivasta et al.**s method update discriminator using history f generated images instead of the latest ones generated. Use a buffer of 50 images
- set $\lambda = 10$ in **TODO: Equation 3**, Adam solver, batchsize 1, trained from scratch with learning rate 0.0002 for the first 100 epochs then decay linearly to 0 in the following 100.

4.2. CyCADA: Cycle Consistent Adversarial Domain Adaptation

see [HTP+17]

4.2.1. Introduction

- synthetic datasets cheaper and more accurate in classification than real ones
- per-pixel label accuracy drops from 93%(real) to 54%(synthetic)
- while translating from synth to real semantic information might be lost (e.g translating line-drawing of a cat to a picture of a dog)
- CyCADA uses cycle consistency and semantic losses
- apply model to digit recognition and semantic segmentation of urban scenes across domains.
- improves per-pixel accuracy from 54% to 82% on synth-to-real. (**TODO: compared to what?**)
- shows that domain adaptation benefits greatly from cycle-consistent pixel transformations
- adaptation at both pixel and representation level can offer complementary improvements with joint pixel-space and feature adaptation leading to the highest performing model for digit classification tasks

4.2.2. Related Work

- **TODO: cite everything**
- visual domain adaptation introduced along with a pairwise metric transform solution by Seanko et al. 2010
- further popularized by broad study of visual dataset bias (Torralba & Efros, 2011)
- early deep adaptive works focused on feature space alignment through minimizing distance between first or second order feature space statistics of source and target (Tzeng et al., 2014; Long & Wand, 2015)
- further improved thorugh use of domain adversarial objectives whereby a domain classifier is trained to distinguish between source and target representations while domain representation is learned so as to maximize error of domain classifier

- representation optimized by using standard minimax objective (Ganin & Lempitsky, 2015)
- symmetric confusion objective (Tzeng et al., 2015)
- inverted label objective (Tzeng et al., 2017)
- each related to GAN (Goodfellow et al., 2014) and followup trianing procedures for these networks (Salimans et al., 2016b; Arjovsky et al., 2017)
- these feature-space adaptation methods focus on modifications to the discriminative representation space. Other recent methods hav sought adaptation in the pixel-space using various generative approaches
- one advantage of pixel-space adaptation: result may be more human interpretable, since an image from on domain can now be visualized in a new domain
- CoGANs (Liu & Tuzel, 2016b) jointly learn source and target representation through explicit weight sharing of certain layers, source and target have unique gen adv objective
- Ghifary et al. 2016 use an additional reconstruction objective in target domain to envourage alignment in the unsupervised adaptation setting
- another approach: directly convert target image into a source style image (or vise versa), largely based on GANs (cite Goodfellow..)
- successfully applied GANs to various applications such as image generation (Denton et al., 2015; Radford et al., 2015; Zhao et al., 2016), image editing (Zhu et al., 2016) and feature learning (Salimans et al., 2016a; Donahue et al., 2017). Recent work (Isola et al., 2016; Sangkloy et al., 2016; Karacan et al., 2016) adopt conditional GANs (Mirza & Osindero, 2014) for these image-to-image translation problems (Isola et al., 2016), but require input-output image pairs for training, which is in general not available in domain adaptation problems
- no training pairs: Yoo et al. 2016 learns source to target encoder-decoder along with a generative adversarial objective on reconstruction which is applied for predicting clothing people are wearing
- Domain Transfer Network (Taigman et al. 2017b) trains generator to transform a source image into a target image by enforccing consistency in embedding space
- Shrivastava et al. 2017 instead use L1 reconstruction loss to force generated target images to be similar to original source images. works well for limited domain shifts where domains are similar in pixel-space, but can be too limiting for setting with larger domain shifts

- Bousmalis et al. 2017b use a content similarity loss to ensure the generated target image is similar to original source image; however this requires prior knowledge about which parts of the image stay the same across domains (e.g. foreground)
- cycada method does not require pre-defining what content is shared between domains and instead simply translates images back to their original domains while ensuring that the remain identical to their original version
- BiGAN (Donahue et al., 2017) and ALI (Dumoulin et al., 2016) take an approach of simultaneously learning the transofmrations between pixel and latent space.
- CycleGAN (Zhu et al., 2017) produced compelling image translation results such as generating photorealistic images from impressionism paintings or transfomring horses into zebras at high resolution using cycle-consistency loss
- this loss was simultaneously proposed by Yi et al. 2017 and Kim et al. 2017 to great effect as well
- adaptation across weather conditions in simple road scenes was first studied by Levinkov & and Fritz 2013
- convolutional domain adversarial based approach was proposed for more general drive cam scenes and for adaptation from simulated to real environments (Hoffmann et al., 2016)
- Ros et al. 2016b learns a multi-source model through concatenating all available labeled data and learning a single large model and then transfers to a sparsely labeled target domain through distillation (Hinton et al., 2015)
- Chen et al. 2017 use an adversarial objective to align both global and class-specific statistics, while mining additional temporal data from street view datasets to learn static object prior
- Zhang et al. 2017 instead perform segmentation adaptation by aligning label distributions both globally and across superpixels in an image

4.2.3. Cycle-consistent adversarial domain adaptation

- consider problem of unsupervised adaptation
- provided source data X_S, source labels Y_S, and target data X_T, but no target labels
- goal: learn a model f that can correctly predict label for target data X_T
- begin by learning source model f_S that can perform the task on the source data

- for K-way classification with cross-entropy loss:

$$\mathcal{L}_{\text{task}}(f_S, X_S, Y_S) = -\mathbb{E}_{(x_s,y_s)\sim(X_S,Y_S)} \sum_{k=1}^{K} \mathbb{1}_{[k=y_s]} \log\Big(\sigma(f_S^{(k)}(x_s))\Big) \tag{4.1}$$

  where σ denotes softmax function

- learned model f_S will perform well on source data but typically domain shift between source and target domain leads to reduced performance when evaluating on target data
- by mapping samples into common space, the model can learn on source data while still generalizing to target data
- mapping from source to target $G_{S\to T}$ trained to produce target samples that fool adversarial discriminator D_T
- discriminator attempts to classify real target data from source target data. loss function:

$$\mathcal{L}_{\text{GAN}}(G_{S\to T}, D_T, X_T, X_S) = \mathbb{E}_{x_t\sim X_T}\Big[\log D_T(x_t)\Big] \tag{4.2}$$

$$+ \mathbb{E}_{x_s\sim X_S}\Big[\log(1 - D_T(G_{S\to T}(x_s)))\Big] \tag{4.3}$$

- this objective ensures that $G_{S\to T}$, given source samples, produces convincing target samples
- this ability to directly map samples between domains allows to learn target model f_T by minimizing $\mathcal{L}_{\text{task}}(f_T, G_{S\to T}(X_S), Y_S)$
- previous approaches that optimized similar objectives have shown effetive results but in practice can often be unstable and prone to failure
- although GAN loss ensures $G_{S\to T}(x_s)$ for some x_s will resemble data drawn from X_T but there is no way to guarantee $G_{S\to T}(x_s)$ preserves structure or content of original sample x_s
- to encourage source content to be preserved during conversion: cycle-consistency constraint (**TODO: cite the 3 works that proposed it**). Mapping $G_{T\to S}$ trained according to GAN loss $\mathcal{L}_{\text{GAN}}(G_{T\to S}, D_S, X_S, X_T)$
- want $G_{T\to S}(G_{S\to T}(x_s)) \approx x_s$ and $G_{S\to T}(G_{T\to S}(x_t)) \approx x_t$
- done by imposing L1 penalty on reconstruction error (reffered to as cycle-consistency loss):

$$\mathcal{L}_{\text{cyc}}(G_{S\to T}, G_{T\to S}, X_S, X_T) = \mathbb{E}_{x_s\sim X_S}\Big[\|G_{T\to S}(G_{S\to T}(x_s)) - x_s\|_1\Big] \tag{4.4}$$

$$+ \mathbb{E}_{x_t\sim X_T}\Big[\|G_{S\to T}(G_{T\to S}(x_t)) - x_t\|_1\Big] \tag{4.5}$$

- also explicitly encourage high semantic consistency before and after image translation
- pretrain source task model f_S, fixing weights and using it as a noisy labeler by which an image to be classified in the same way after translation as it was before translation according to this classifier is encouraged
- define fixed classifier f, predicted label for given input X: $p(f, X) = \arg\max(f(X))$
- semantic consistency before and after image translation:

$$\mathcal{L}_{\text{sem}}(G_{S\to T}, G_{T\to S}, X_S, X_T, f_S) = \mathcal{L}_{\text{task}}(f_S, G_{T\to S}(X_T), p(f_S, X_T)) \tag{4.6}$$
$$+ \mathcal{L}_{\text{task}}(f_S, G_{S\to T}(X_S), p(f_S, X_S)) \tag{4.7}$$

- can be viewed analogously to content loss in style transfer (Gatys et al., 2016) or in pixel adaptation (Taigman et al., 2017a), where shared content to preserve is dictated by the source task model f_S
- could also consider a feature-level method which discriminates between the features or semantics from two image sets as viewed under a task network → additional feature level GAN loss:

$$\mathcal{L}_{\text{GAN}}(f_T, D_{\text{feat}}, f_S(G_{S\to T}(X_S)), X_T) \tag{4.8}$$

- together form complete objective:

$$\mathcal{L}_{\text{CyCADA}}(f_T, X_S, X_T, Y_S, G_{S\to T}, G_{T\to S}, D_S, D_T) \tag{4.9}$$
$$= \mathcal{L}_{\text{task}}(f_T, G_{S\to T}(X_S), Y_S) \tag{4.10}$$
$$+ \mathcal{L}_{\text{GAN}}(G_{S\to T}, D_T, X_T, X_S) + \mathcal{L}_{\text{GAN}}(G_{T\to S}, D_S, X_S, X_T) \tag{4.11}$$
$$+ \mathcal{L}_{\text{GAN}}(f_T, D_{\text{feat}}, f_S(G_{S\to T}(X_S)), X_T) \tag{4.12}$$
$$+ \mathcal{L}_{\text{cyc}}(G_{S\to T}, G_{T\to S}, X_S, X_T) + \mathcal{L}_{\text{sem}}(G_{S\to T}, G_{T\to S}, X_S, X_T, f_S) \tag{4.13}$$

- ultimately corresponds to solving for a target model f_T according to the optimization problem

$$f_T^* = \arg\min_{f_T} \min_{\substack{G_{T\to S}\\ G_{S\to T}}} \max_{D_S, D_T} \mathcal{L}_{\text{CyCADA}}(f_T, X_S, X_T, Y_S, G_{S\to T}, G_{T\to S}, D_S, D_T) \tag{4.14}$$

5. Conclusion

To conclude...

A. Blub

Bibliography

[DBP+16] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[DKD16] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[HTP+17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *CoRR*, abs/1711.03213, 2017.

[SKFD10] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 213–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[SUHS17] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *CoRR*, abs/1711.01575, 2017.

[THSD17] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *CoRR*, abs/1702.05464, 2017.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.