

# Unsupervised Domain Adoption by Backpropogation

Патакин Николай

23 января 2020 г.

# Domain adoption

Существует проблема разнородности train и test данных:

- Данные могут быть получены из различных источников:



digital SLR camera



low-cost camera, flash



amazon.com



consumer images

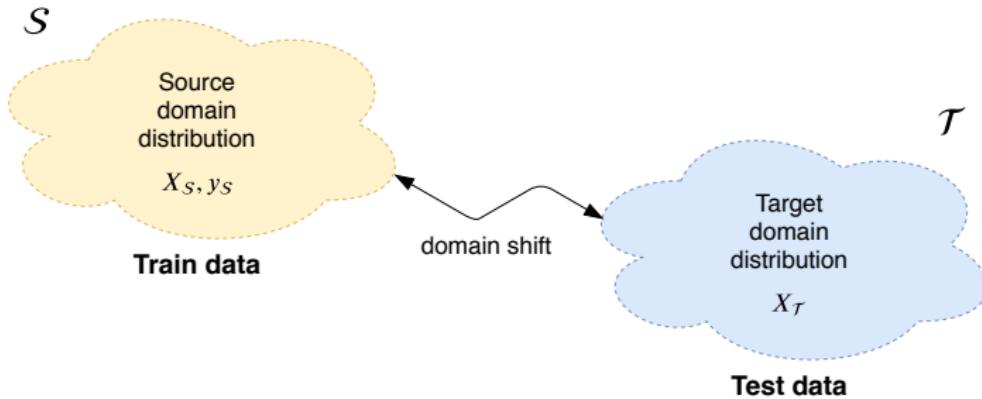


# Domain adoption

- Могут использоваться различные методики сбора данных  
(Для некоторых проблематично получить разметку, но имеют точное распределение, другие - неточное распределение, но дешевая разметка)
- Синтетические данные



# Domain adoption



В зависимости от сложности ситуации:

- Данных Target достаточно  $\Rightarrow$  выполняем fine-tuning
- Supervised DA - Target размечен, но данных мало
- Semi-supervised DA - часть данных Target размечена, но есть неразмеченные данные
- **Unsupervised DA** - полностью отсутствует разметка для Target

- Поиск геометрического преобразования распределений  $\mathcal{S} \rightarrow \mathcal{T}$ :
  - Kernel-преобразования данных и выполнения PCA, SVD разложений (Gong & Grauman 2012 - GFK, Fernando 2013 - SA)
  - Поиск преобразования с помощью landmark selection (Gong & Grauman 2013)
  - Постепенное преобразование доменов с помощью deep autoencoders (Chopra & Gopalan 2013)

- Поиск геометрического преобразования распределений  $\mathcal{S} \rightarrow \mathcal{T}$ :
  - Kernel-преобразования данных и выполнения PCA, SVD разложений (Gong & Grauman 2012 - GFK, Fernando 2013 - SA)
  - Поиск преобразования с помощью landmark selection (Gong & Grauman 2013)
  - Постепенное преобразование доменов с помощью deep autoencoders (Chopra & Gopalan 2013)
- Построение domain-invariant признакового описания объектов

# Формализация задачи

Имеется пространство объектов  $X$ , целевых меток  $Y$ .

$\mathcal{S}(x, y)$  – Source domain distribution,  $x \in X, y \in Y$

$\mathcal{T}(x, y)$  – Target domain distribution

$G_f(x; \theta_f)$  – Feature encoder

Требуется добиться инвариантности относительно домена распределений признаков:

$$\mathcal{S}(f) = \{G_f(x; \theta_f) \mid x \sim \mathcal{S}(x)\}$$

$$\mathcal{T}(f) = \{G_f(x; \theta_f) \mid x \sim \mathcal{T}(x)\}$$

Как измерять близость распределений?

- KL-divergence (Gong & Grauman 2012)
- Kernel reproducing Hilbert space ( $H\Delta H$ -divergence)
- Tzeng et al. 2014:

$$MMD = \left\| \frac{1}{|X_s|} \sum_i \varphi(x_{s,i}) - \frac{1}{|X_T|} \sum_j \varphi(x_{t,j}) \right\|_2^2$$

# Предлагаемый подход

Построение признаков  $G_f(x; \theta)$

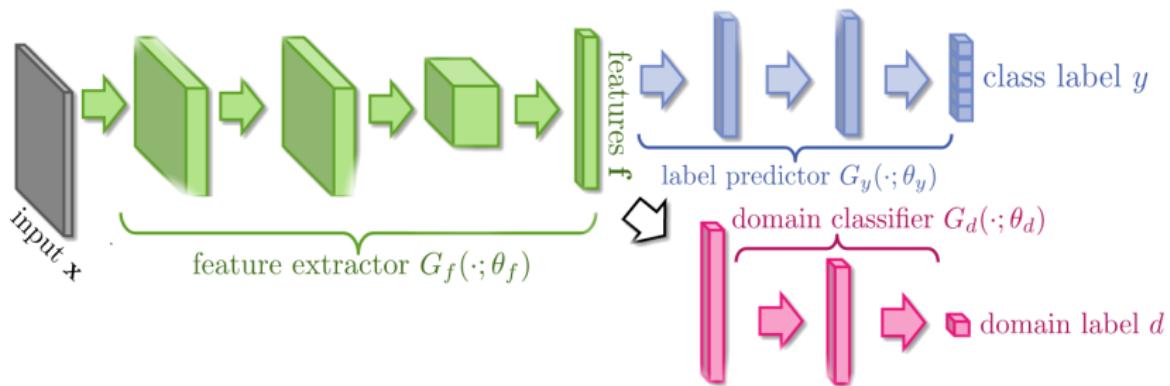
(а) дискриминативных

(б) доменно-инвариантных



label predictor

domain classifier



# Оптимизируемый функционал

$$E(\theta_f, \theta_y, \theta_d) = \sum_{\substack{i=1..N \\ d_i=0}} L_y(G_y(G_f(x_i; \theta_f); \theta_y), y_i) - \lambda \sum_{i=1..N} L_d(G_d(G_f(x_i; \theta_f); \theta_d), d_i)$$

Обучаем feature extractor и label predictor добиваясь дискриминативности и инвариантности:

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg \min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \theta_d)$$

При этом пытаемся усилить дискриминатор:

$$\hat{\theta}_d = \arg \max_{\theta_d} E(\theta_f, \theta_y, \theta_d)$$

Улучшаем классификатор и дискриминатор:

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y}$$

$$\theta_d \leftarrow \theta_d - \mu \frac{\partial L_d^i}{\partial \theta_d}$$

# Оптимизируемый функционал

Улучшаем классификатор и дискриминатор:

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y}$$

$$\theta_d \leftarrow \theta_d - \mu \frac{\partial L_d^i}{\partial \theta_d}$$

Тогда, feature extractor обновлялся бы:

$$\theta_f \leftarrow \theta_f - \mu \left( \frac{\partial L_y^i}{\partial \theta_y} + \frac{\partial L_d^i}{\partial \theta_d} \right)$$

# Оптимизируемый функционал

Улучшаем классификатор и дискриминатор:

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y}$$

$$\theta_d \leftarrow \theta_d - \mu \frac{\partial L_d^i}{\partial \theta_d}$$

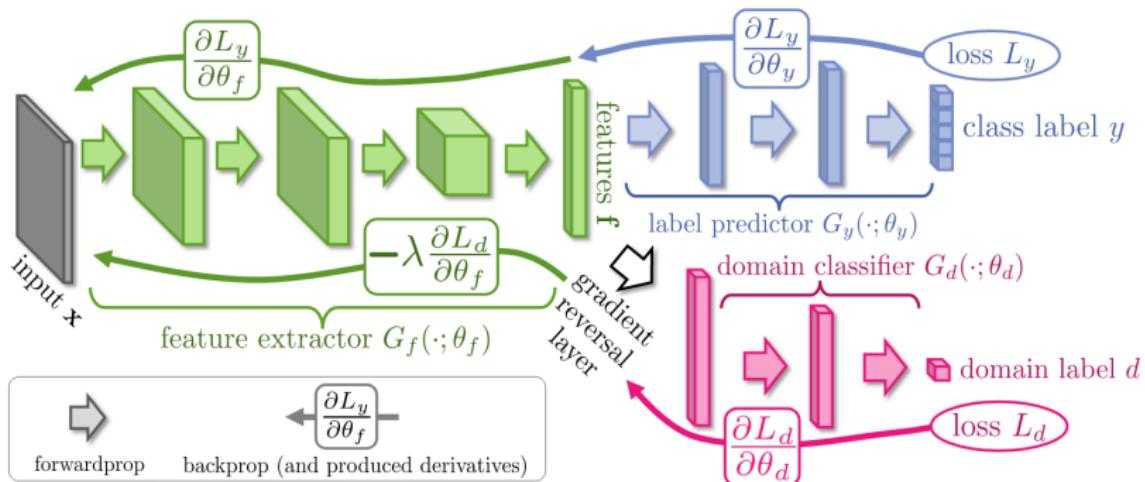
Тогда, feature extractor обновлялся бы:

$$\theta_f \leftarrow \theta_f - \mu \left( \frac{\partial L_y^i}{\partial \theta_y} \cancel{+} \frac{\partial L_d^i}{\partial \theta_d} \right)$$

$$\theta_f \leftarrow \theta_f - \mu \left( \frac{\partial L_y^i}{\partial \theta_y} - \lambda \frac{\partial L_d^i}{\partial \theta_d} \right)$$

# Gradient reversal layer

GRL. Forward:  $R_\lambda(x) = x$ , Backward:  $\frac{\partial R_\lambda}{\partial x} = -\lambda \mathbf{I}$



- Оптимизация при помощи SGD(momentum=0.9).
- Использование мини-батча размером 128. 50% - source domain, 50% - target domain
- На начальных этапах дискриминатор имеет шумный выход, используется плавное увеличение  $\lambda$ :

$$\lambda_p = \frac{2}{1 + \exp(-\gamma p)} - 1$$

- Оптимизация гиперпараметров по валидационной выборке source domain и ошибке классификации домена

# Используемые датасеты

	MNIST	SYN NUMBERS
SOURCE		
TARGET		
	MNIST-M	SVHN
SVHN		
MNIST		SYN SIGNS
GTSTB		

# Результаты

METHOD	SOURCE	MNIST	SYN NUMBERS	SVHN	SYN SIGNS
	TARGET	MNIST-M	SVHN	MNIST	GTSRB
SOURCE ONLY		.5749	.8665	.5919	.7400
SA (FERNANDO ET AL., 2013)		.6078 (7.9%)	.8672 (1.3%)	.6157 (5.9%)	.7635 (9.1%)
PROPOSED APPROACH		<b>.8149 (57.9%)</b>	<b>.9048 (66.1%)</b>	<b>.7107 (29.3%)</b>	<b>.8866 (56.7%)</b>
TRAIN ON TARGET		.9891	.9244	.9951	.9987

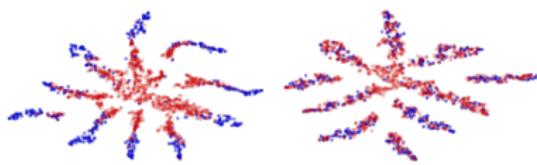
MNIST → MNIST-M: top feature extractor layer



(a) Non-adapted

(b) Adapted

SYN NUMBERS → SVHN: last hidden layer of the label predictor



(a) Non-adapted

(b) Adapted

# Office dataset



# Office dataset

METHOD	SOURCE	AMAZON	DSLR	WEBCAM
	TARGET	WEBCAM	WEBCAM	DSLR
GFK(PLS, PCA) (GONG ET AL., 2012)		.464 ± .005	.613 ± .004	.663 ± .004
SA (FERNANDO ET AL., 2013)		.450	.648	.699
DA-NBNN (TOMMASI & CAPUTO, 2013)		.528 ± .037	.766 ± .017	.762 ± .025
DLID (S. CHOPRA & GOPALAN, 2013)		.519	.782	.899
DECAF <sub>6</sub> SOURCE ONLY (DONAHUE ET AL., 2014)		.522 ± .017	.915 ± .015	–
DANN (GHIFARY ET AL., 2014)		.536 ± .002	.712 ± .000	.835 ± .000
DDC (TZENG ET AL., 2014)		.594 ± .008	.925 ± .003	.917 ± .008
PROPOSED APPROACH		<b>.673 ± .017</b>	<b>.940 ± .008</b>	<b>.937 ± .010</b>

Идея построения доменно-инвариантных признаков в совокупности с измерением меры адаптированности доменов посредством loss-функции дискриминатора показывает отличный результат

- Метод очень просто адаптируется не только к различным моделям, но и к различным задачам
- Использование моделей глубинного обучения с end-to-end процедурой обучения
- Прост в реализации и интуитивен
- Показывает значительно лучшее качество по сравнению с предыдущими подходами

- В каких ситуациях возникает необходимость в адаптации доменов? Приведите примеры. В чем заключается классический подход к этой задаче и в чем разница с предлагаемым авторами Domain Adoption by Backpropogation
- Запишите оптимизируемый функционал в Domain Adoption by Backpropogation, поясните все компоненты входящие в его состав
- Почему невозможна оптимизация функционала Domain Adoption by Backpropogation при помощи SGD? Сформируйте понятие Gradient Reversal Layer (GRL), объясните принцип его действия