Вариационное моделирование правдоподобия с триплетными ограничениями в задачах информационного поиска

М.В. Кузнецова

Интеллектуализация обработки информации - 2020
Московский физико-технический институт

Постановка задачи

В машинном обучении есть обширный ряд задач, где объекты представленны разными модальностями, например:

- соответствующие друг другу тексты на разных языках;
- дневные и ночные фотографии одной местности;
- фотографии одной местности с разных точек;
- фотографии двойняшек разного пола;
- записи разных исполнений музыкального произведения и др.

Есть множества (домены) $\mathrm{X} = \{\mathrm{x}_a\}_{a=1}^{P} \subset \mathrm{X}_{all}$ и $\mathrm{Y} = \{\mathrm{y}_b\}_{b=1}^{Q} \subset \mathrm{Y}_{all}$ всех объектов одного и того же типа. Выборка $(\mathrm{X}, \mathrm{Y}) = \{(\mathrm{x}, \mathrm{y})_i\}_{i=1}^{N}$, состоит из N пар объектов (элементов доменов). Предполагается, что каждая пара $(\mathrm{x}, \mathrm{y})_i$ отражает разные модальности одного и того же объекта.

Постановка задачи

На декартовом произведении X_{all}, Y_{all} определена функция sim_{all}:

$$sim_{all} : (x_{c_{all}}, y_{d_{all}})_{x_{c_{all}} \in X_{all}, y_{d_{all}} \in Y_{all}} \to \{0, 1\},$$

которая паре $(x_{c_{all}}, y_{d_{all}})$ ставит в соответствие 0 или 1 по следующему правилу:

$$sim_{all}(x_{c_{all}}, y_{d_{all}}) = \begin{cases} 1, & \text{когда } x_{c_{all}} \text{ и } y_{d_{all}} \text{ являются разными модальностями} \\ & \text{одного и того же объекта;} \\ 0, & \text{иначе.} \end{cases}$$

На данной наблюдаемой выборке (X, Y) задана функция sim с частично известными значениями:

$$sim(x_c, y_d) = \begin{cases} 1, & \text{когда } x_c \text{ и } y_d \text{ являются разными модальностями} \\ & \text{одного и того же объекта;} \\ 0, & \text{нет информации о соответствии объектов в паре.} \end{cases}$$

Известно, что на некоторой небольшой части пар $\epsilon << N$ sim может принимать ошибочное значение, т.е некоторые пары из (X, Y) сопоставлены некорректно. Требуется найти отображение $\hat{f}$ по выборке (X, Y), такое, что на $\mathcal{D}_{test}$ достигается минимум некоторой функции потерь l:

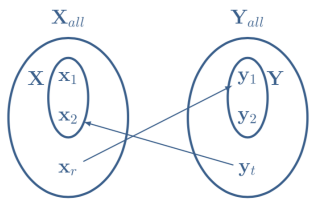
$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x_{c_{test}}, y_{d_{test}}) \in \mathcal{D}_{test}} l(f(x_{c_{test}}, y_{d_{test}}), r), \tag{1}$$

$r \in \{0, 1\}$, $\mathcal{F}$ —заданное семейство моделей.

Постановка задачи. Примеры подзадач

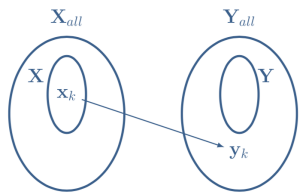
Требуется построить алгоритм для:
информационного поиска: по заданному x найти y, удовлетворяющий некоторому критерию, и наоборот (прямой и обратный инф. поиск).

$$f_{search\ direct} : X_{all} \to Y, f_{search\ reverse} : Y_{all} \to X$$



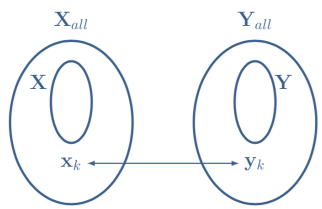
генерации объекта $\hat{y}$ из по заданному x и наоборот.

$$f_{gen} : X \to \hat{Y} \subseteq Y_{all} : \forall\ x\ \in\ X\ \exists\ \hat{y}\ \in\ \hat{Y} : sim_{all}(x, \hat{y}) = 1$$



двунаправленной генерации условно-реальных пар $(\hat{x}, \hat{y})$.

$$f_{bigen} : X, Y \to \hat{X}, \hat{Y} \subseteq X_{all}, Y_{all} : \forall \hat{x}\ \in\ \hat{X},\ \forall \hat{y}\ \in\ \hat{Y}\ sim_{all}(\hat{x}, \hat{y}) = 1$$

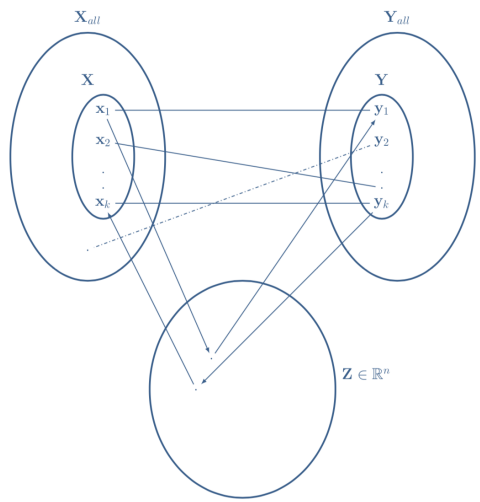


Проблема

1. Объекты доменов имеют сложную структуру, разную для каждой задачи.
2. Соответствие между объектами доменов нечеткое:
   - 2.1 объекты в паре могут быть сопоставлены некорректно;
   - 2.2 соотношение между объектами доменов, вообще говоря, может быть многие к многим.

Предлагается

Ввести промежуточное пространство (пространство оценок) Z для отображения туда объектов из X_{all} и Y_{all} и решения перечисленных задач в $\mathbb{R}^n$. Предлагается строить отображение в пространство оценок, используя вариационные методы.



Формальная постановка задачи (один домен)

- Задана выборка $\mathrm{X} = \{\mathrm{x}\}_{i=1}^{N}$, состоящая из N *i.i.d.* объектов
- Предположим что объекты выборки X сгенерированы при условии скрытой случайной величины $\mathrm{z} \in \mathbb{R}^n$.

Введем вспомогательное распределение $q_\phi(\mathrm{z}|\mathrm{x})$. Проводится оптимизация вариационной нижней оценки:

$$\log p_\theta(\mathrm{x}) \geq \underbrace{\int q_\phi(\mathrm{z}|\mathrm{x}) \log p_\theta(\mathrm{x}|\mathrm{z}) d\mathrm{z}}_{\text{Ошибка реконструкции}} - \underbrace{KL(q_\phi(\mathrm{z}|\mathrm{x})||p(\mathrm{z}))}_{\text{Штраф}} = \mathcal{L}$$

Решается задача оптимизации нижней вариационной оценки по параметрам θ и ϕ:

$$\mathcal{L} \to \max_{\theta,\phi}$$

C. M. Bishop. Pattern recognition. – 2006

Формальная постановка задачи (два домена)

Для выборки $\{(\mathrm{x},\mathrm{y})_i\}_{i=1}^N$ определим правдоподобие как:

$$p(\mathrm{x},\mathrm{y}) = \int_{\mathrm{z}} p_{\theta_\mathrm{x}}(\mathrm{x}|\mathrm{z}) p_{\theta_\mathrm{y}}(\mathrm{y}|\mathrm{z}) d\mathrm{z},$$

где z — скрытое представление пары x, y.

Для оптимизации правдоподобия введем вспомогательное распределения $q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})$. Предполагается следующий генеративный процесс:

- генерируем z из апостериорного распределения $p(\mathrm{z}) = \mathcal{N}(0, \mathrm{I})$,
- x генерируем из апостериорного распределения $p_{\theta_\mathrm{x}}(\mathrm{x}|\mathrm{z})$,
- y генерируем из апостериорного распределения $p_{\theta_\mathrm{y}}(\mathrm{y}|\mathrm{z})$.

Ставится задача максимизации логарифма правдоподобия:

$$\mathcal{L}_{VBA} \to \max_{\theta,\phi}$$

Suzuki, M., Nakayama, K., Matsuo, Y. Joint multimodal learning with deep generative models. – 2016.

Вариационный автокодировщик для двух доменов

Логарифм правдоподобия Вариационного автокодировщика [1] для двух доменов (VBA) представим в виде:

$$\mathcal{L}_{VBA} = \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} \log \frac{p_\theta(\mathrm{x},\mathrm{y},\mathrm{z})}{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} =$$

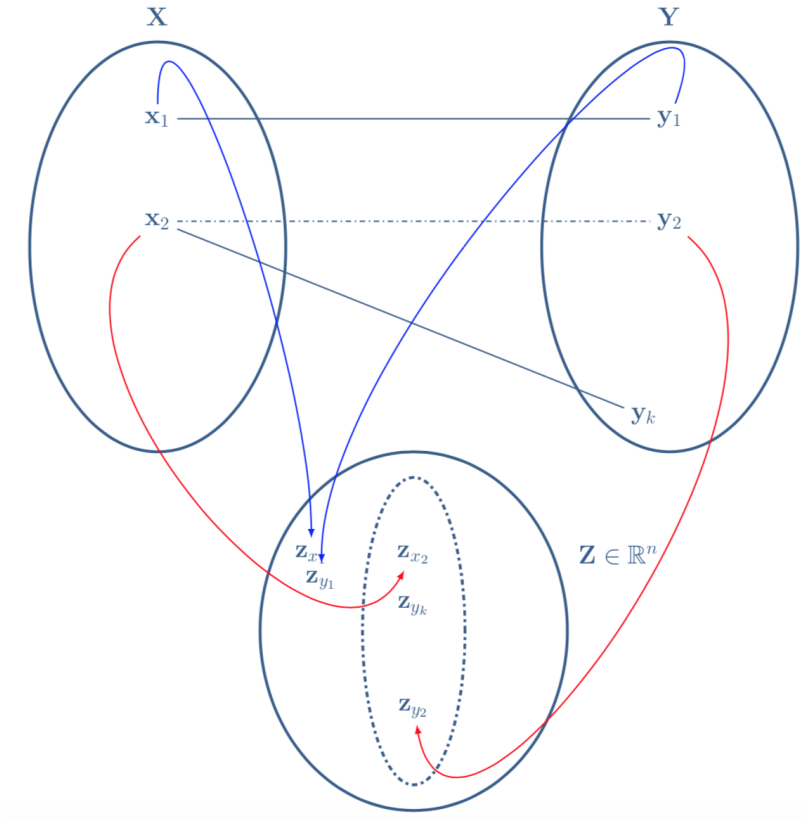
$$= \underbrace{-KL\big(q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y}) \parallel p(\mathrm{z})\big)}_{\text{Штраф}} + \underbrace{\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})}\big[\log p_{\theta_\mathrm{y}}(\mathrm{y}|\mathrm{z}) + \log p_{\theta_\mathrm{x}}(\mathrm{x}|\mathrm{z})\big]}_{\text{Ошибка реконструкции}}$$

где $q_\phi(\mathrm{z}_x|\mathrm{x},\mathrm{y})$ — энкодер, $p_{\theta_\mathrm{x}}(\mathrm{x}|\mathrm{z})$ и $p_{\theta_\mathrm{y}}(\mathrm{y}|\mathrm{z})$ — декодеры, параметризуемые нейронными сетями.

[1] Kingma, D. P., Welling, M. Auto-encoding variational bayes. – 2013.

Проблема

При наличии шума в данных нужно научить *VBA*-модель выявлять корректные соответствия для рядом расположенных объектов.
Нужно внести в модель штраф за несоответствие пар.



Триплетные ограничения

Выборка состоит из пар $\{(x, y)_i\}_{i=1}^N$, где предполагается, что объекты каждой пары (x, y) близки к друг другу в некотором смысле. Определим верный триплет следующим образом:

$$\mathcal{T} = \{t_{x,y,k}\} = \{(z_x, z_y, z_k) : d(z_x, z_y) < d(z_x, z_k)\},$$

где

- d — некоторая функция близости;
- $z_k \in \{Z_X \cup Z_Y\}$, Z_X и Z_Y — скрытые представления всех объектов выборки (X, Y);
- z_k — некоторое преобразование объекта k;
- z_x, z_y — скрытые представления объектов пары (x, y).

Общее правдоподобие модели с триплетными ограничениями

Определим правдоподобие триплетного ограничения как:

$$p(t_{x,y,k} \in \mathcal{T} | x, y, k) = \int_{\mathrm{z}} p(\mathrm{t}_{x,y,k} | \mathrm{z}_x, \mathrm{z}_y, \mathrm{z}_k) p(\mathrm{z}_x) p(\mathrm{z}_y) p(\mathrm{z}_k) d\mathrm{z}_x d\mathrm{z}_y d\mathrm{z}_k,$$

Функцию правдоподобия моделируется распределением Бернулли, параметризованным функцией softmax, аналогично с Karaletsos [2] для случая одного домена:

$$p(t_{x,y,k} \in \mathcal{T} | x, y, k) = Ber(t_{x,y,k} \in \mathcal{T} | x, y, k) = \frac{e^{-d(\mathrm{z}_x, \mathrm{z}_y)}}{e^{-d(\mathrm{z}_x, \mathrm{z}_y)} + e^{-d(\mathrm{z}_x, \mathrm{z}_k)}}.$$

[2] T. Karaletsos et al., Bayesian representation learning with oracle constraints //International Conference of Learning Representations – 2016.

Общее правдоподобие модели с триплетными ограничениями

Для выборки $\{(\mathrm{x}, \mathrm{y})_i\}_{i=1}^N$ правдоподобие с триплетным ограничением определяется как:

$$p(\mathrm{x}, \mathrm{y}, \mathrm{t}) = \int_{\mathrm{z}} p_{\theta_\mathrm{x}}(\mathrm{x}|\mathrm{z}_x) p_{\theta_\mathrm{y}}(\mathrm{y}|\mathrm{z}_y) p(t_{x,y,k}|\mathrm{z}_x, \mathrm{z}_y, \mathrm{z}_k) d\mathrm{z}.$$

Ставится задача максимизации логарифма правдоподобия:

$$\mathcal{L}_{VBTA}(\mathrm{x}, \mathrm{y}, \mathrm{t}) \to \max_{\theta_\mathrm{x}, \theta_\mathrm{y}, \phi_\mathrm{x}, \phi_\mathrm{y}}$$

Вариационный автокодировщик с триплетными ограничениями

Теорема

Обозначим $\mathrm{z}_x, \mathrm{z}_y, \mathrm{z}_k$ как $\mathrm{z}_{x,y,k}$. Логарифм правдоподобия Вариационного автокодировщика с триплетными ограничениями (*VBTA*) представим в виде:

$$\mathcal{L}_{VBTA} = \mathbb{E}_{q_\phi(\mathrm{z}_{x,y,k}|\mathrm{x},\mathrm{y},\mathrm{t})} \log \frac{p_\theta(\mathrm{x},\mathrm{y},\mathrm{t},\mathrm{z}_{x,y,k})}{q_\phi(\mathrm{z}_{x,y,k}|\mathrm{x},\mathrm{y},\mathrm{t})} =$$

$$= \underbrace{-KL\big(q_\phi(\mathrm{z}_{x,y,k}|\mathrm{x},\mathrm{y},\mathrm{t}) \parallel p(\mathrm{z})\big)}_{\text{Штраф}} +$$

$$+ \underbrace{\Big[\mathbb{E}_{q_{\phi_\mathrm{x}}(\mathrm{z}_x|\mathrm{x})}\big[\log p_{\theta_\mathrm{x}}(\mathrm{x}|\mathrm{z}_x)\big] + \mathbb{E}_{q_{\phi_\mathrm{y}}(\mathrm{z}_y|\mathrm{y})}\big[\log p_{\theta_\mathrm{y}}(\mathrm{y}|\mathrm{z}_y)\big]\Big]}_{\text{Ошибка реконструкции}} +$$

$$+ \underbrace{\mathbb{E}_{q_\phi(\mathrm{z}_{x,y,k}|\mathrm{x},\mathrm{y},\mathrm{t})}\big[\log p(t_{x,y,k}|\mathrm{z}_x,\mathrm{z}_y,\mathrm{z}_k)\big]}_{\text{Ошибка триплетов}}$$

Функция влияния

Пусть G — эмпирическая (выборочная) функция распределения выборки X:

$$G(\mathrm{x}) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathrm{x}_i \leqslant \mathrm{x}\}}, \text{ где } 1_{\{\mathrm{x}_i \leqslant \mathrm{x}\}} \text{ — индикаторная функция, далее } \triangle_{\mathrm{x}^o}\{\mathrm{x}\}.$$

Пусть $G_{\epsilon,\mathrm{x}_o}$ — "зашумленная" версия эмпирической функции распределения в точке x^o, ϵ — уровень шума:

$$G_{\epsilon,\mathrm{x}_o} = (1-\epsilon)G(\mathrm{x}) + \epsilon\triangle_{\mathrm{x}_o}(\mathrm{x}).$$

Для статистики T и распределения G функция влияния (IF) в точке x_o определяется следующим образом [3]:

$$IF(\mathrm{x}_o, T, G) = \frac{\partial}{\partial\epsilon} T(G_{\epsilon,\mathrm{x}_o}(\mathrm{x}))\Big|_{\epsilon=0} = \lim_{\epsilon\to 0}\frac{T(G_{\epsilon,\mathrm{x}_o}(\mathrm{x})) - T(G(\mathrm{x}))}{\epsilon}.$$

[3] Hampel F. R. et al. Robust statistics: the approach based on influence functions, 2011.

Функция влияния

Теорема

Пусть $G_\epsilon(\mathrm{x},\mathrm{y}) = (1-\epsilon)G(\mathrm{x},\mathrm{y}) + \epsilon \triangle_{(\mathrm{x},\mathrm{y})^o}(\mathrm{x},\mathrm{y})$ — эмпирическая функция распределения выборки, тогда функция влияния VBTA-модели определяется как:

$$\left(\frac{\partial^2 \mathcal{L}_{VBTA}}{\partial \phi^2}\right)^{-1} \nabla_\phi \mathbb{E}_{q_\phi}\left[KL\Big(q_\phi \parallel p(\mathrm{z})\Big) + N \log(\mathrm{x}^o, \mathrm{y}^o, \mathrm{t}^o|\mathrm{z})\right],$$

где t^o — “шумный” триплет, тройка объектов, содержащих “шумную” пару.

Теорема

Если триплетное правдоподобие моделируется распределением Стьюдента то $sup_{(\mathrm{x},\mathrm{y})^o}|IF((\mathrm{x},\mathrm{y})^o, T, G)|$ для *VBTA*-модели ограничен при условии того, что оптимизация ошибки реконструкции не проводится. Следовательно, *VBTA*-модель является устойчивой к шуму в выборке.

Эксперименты на выборке MNIST

Выборка пар: MNIST и перевернутый MNIST.

Модель	$n = 0$	$n = 10$	$n = 100$	$n = 1000$	All
DiscoGAN	-	-	-	-	15.00
	-	-	-	-	$\pm$0.20
Triple GAN	-	-	63.79	84.93	86.70
	-	-	$\pm$0.85	$\pm$1.63	$\pm$1.52
Δ-GAN	-	-	83.20	88.98	93.34
	-	-	$\pm$1.88	$\pm$1.50	$\pm$1.46
VBTA	18.89	86.57	**90.44**	**90**	**95**
	$\pm$3.59	$\pm$6.34	$\pm$**0.3**	$\pm$**0.26**	$\pm$**0.06**

Точность классификации; n — количество объектов, использованных для формирования триплетных ограничений.

Модель	$< log p(\mathrm{x})$	$< log p(\mathrm{y})$
VAE	-81.13	-81.01
JVMAE	-85.35	-85.44
Proposed	-**80.92**	-**80.91**

Нижняя вариационная оценка.

Эксперимент с шумом в выборке

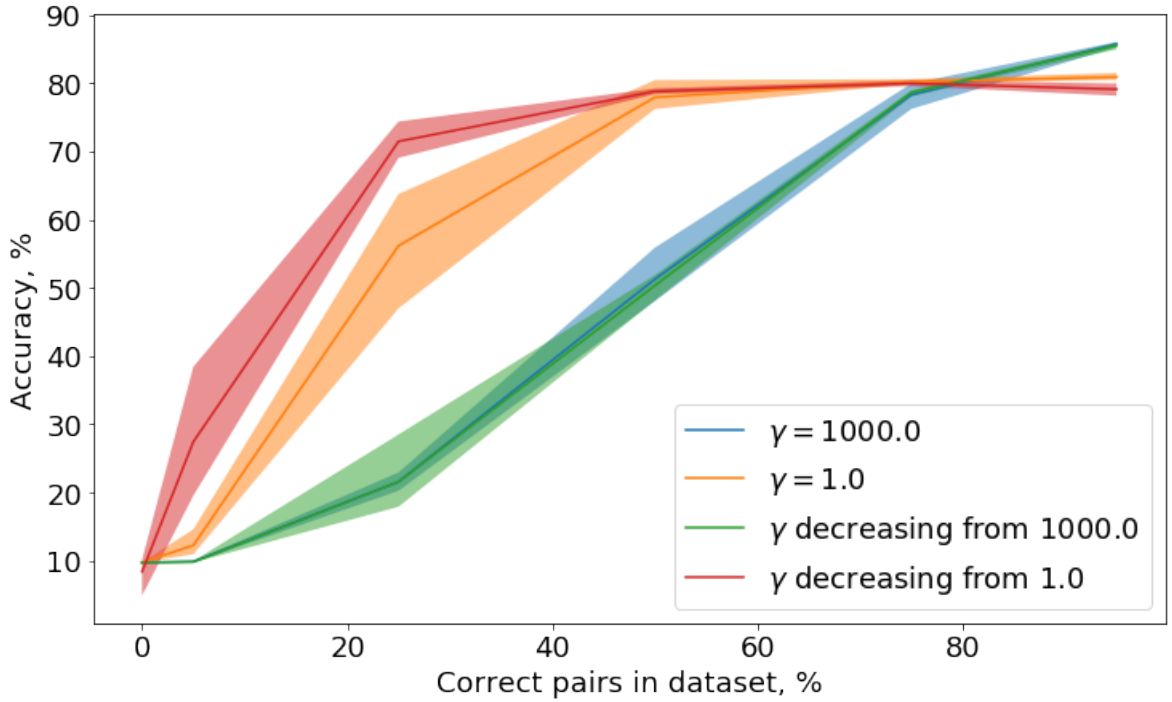
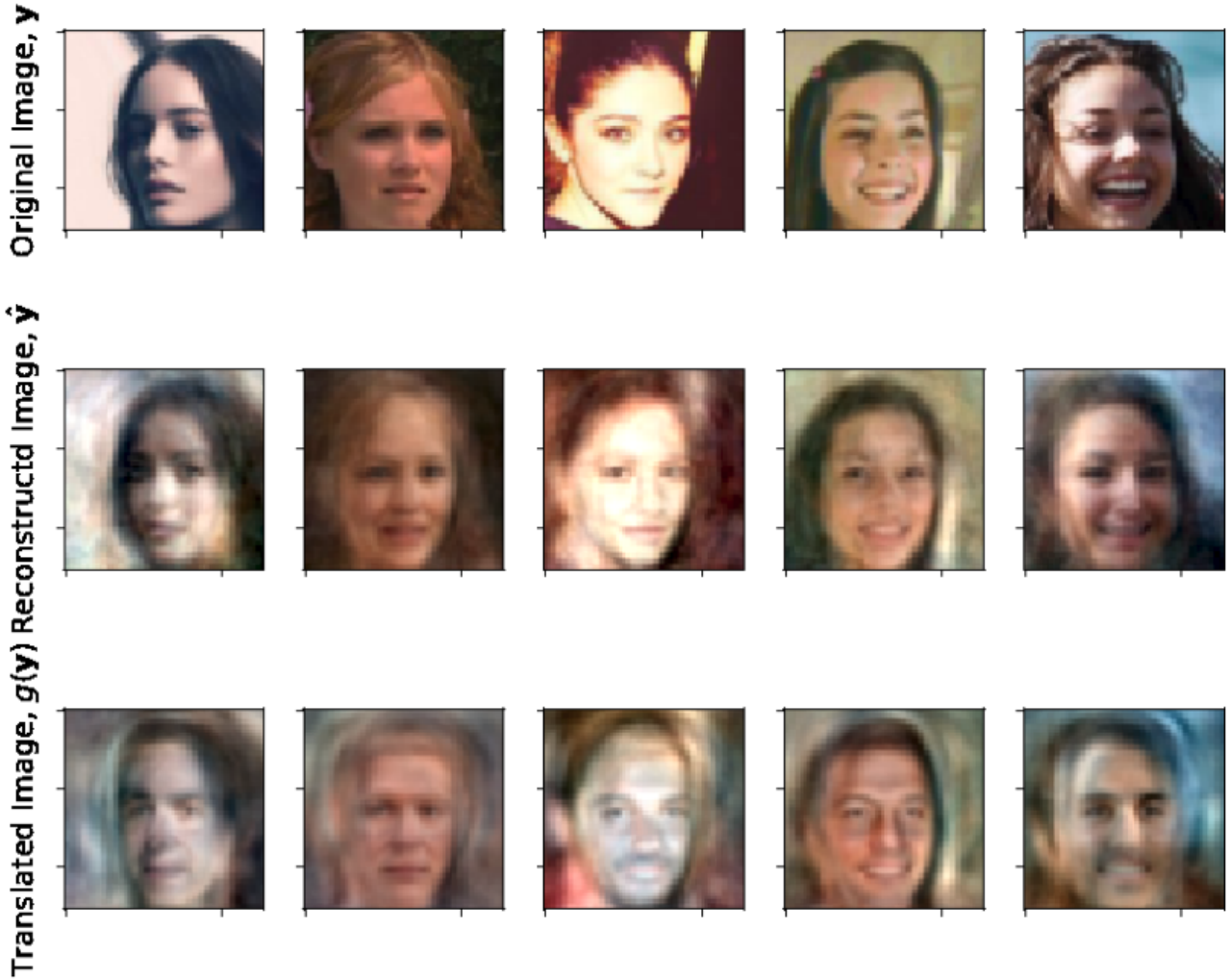


Рис.: Точность VBTA -модели с разным уровнем шума в выборке.

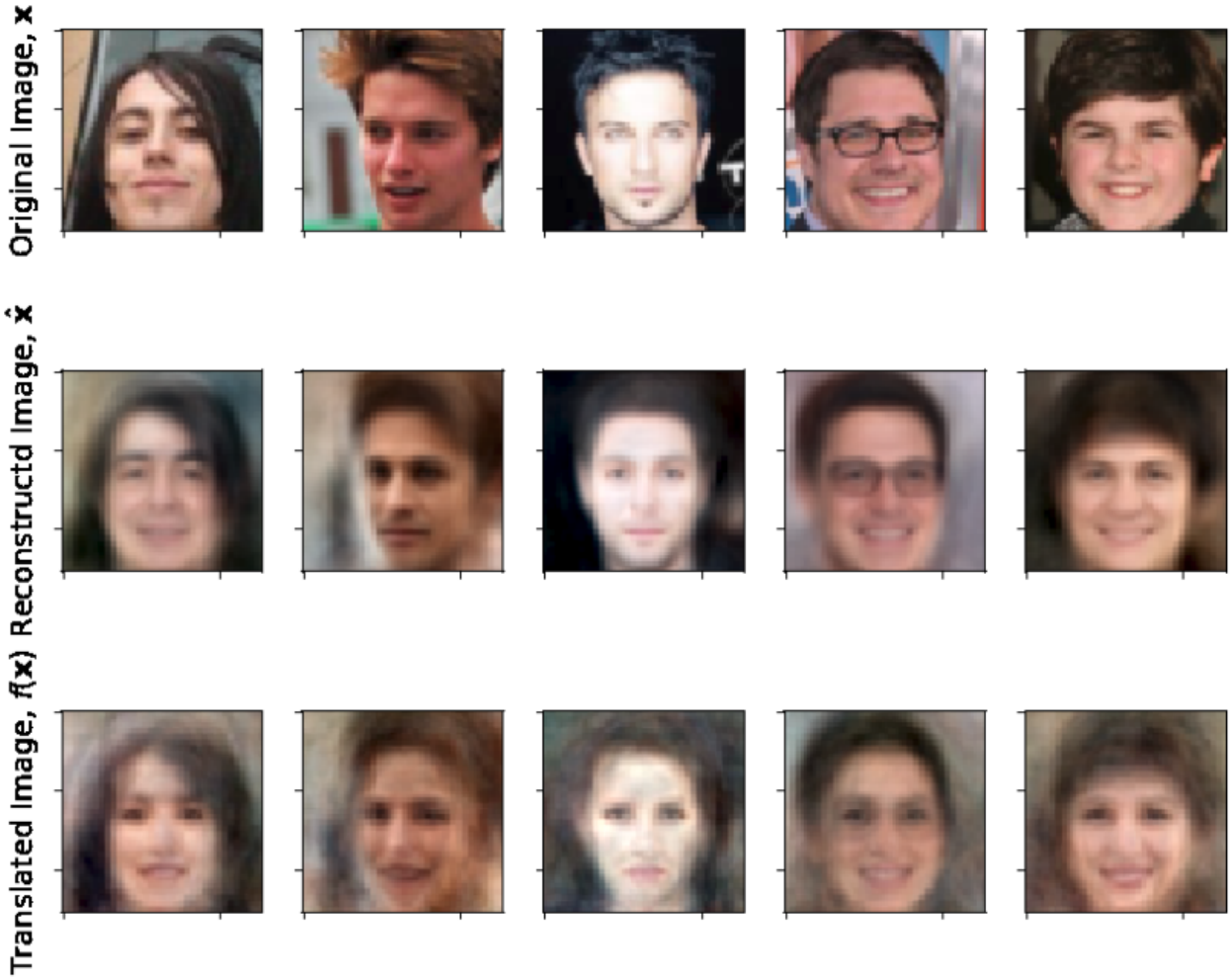
1. $\gamma = 1$;
2. $\gamma = 1000$;
3. линейное понижение γ от 1000 до 0;
4. линейное понижение γ от 1 до 0.

Выборка Celeba — генерация объекта по заданному



Результат генерации объекта по заданному. Первая строка соответствует заданному изображению, вторая — его реконструкции, в третьей строке — сгенерированный объект.

Выборка Celeba — генерация объекта по заданному



Результат генерации объекта по заданному. Первая строка соответствует заданному изображению, вторая — его реконструкции, в третьей строке — сгенерированный объект.

Выборка Celeba

Модель	Accuracy, %
Men reconstruction	77%
Women reconstruction	98%
Men2Women	95%
Women2Men	89%
Men generation	73%
Women generation	97%

CelebA — точность классификации.

Модель	Accuracy %
Zhao et al. (2018)	97%
Yu et al.(2018)	90%
Choi et al.(2017)	98%
Wang et al.(2018)	82%
VBTA (генерация)	92%
VBTA (двун. генерация)	85%

Celeba — сравнение с другими подходами.

Кросс-языковая классификация документов

Дан классификатор, обученный на текстах на языке A, нужно решить задачу классификации для языка B. Для решения этой задачи будем использовать *VBTA*-модель для построения двуязычных текстовых представлений. Решается задача многоклассовой классификации, где меткой класса каждого документа является одна из четырех тем.

Model	$en \rightarrow de$	$de \rightarrow en$
Klementiev et al., 2012	77.6	71.1
Gouws et al., 2015	86.5	75.0
Chandar et al., 2014	91.8	74.2
Wei and Deng, 2017	92.7	80.4
Su et al., 2018	91.3	77.8
VBTA	**94.3**	**82.8**

Точность классификации

Результаты

- Предложена новая модель вариационного автокодировщика, моделирующего совместное правдоподобие данных разных доменов.
- Предложена модель, основанная на триплетных ограничениях, позволяющая снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач.
- Доказана устойчивость модели.
- Проведены эксперименты на модельных и реальных данных.