Текстово-визуальный кросс-модальный поиск без использования параллельных данных

Григорий Бартош

Научный руководитель: к.т.н. П. И. Браславский

Научный консультант: А. А. Шпильман

Санкт-Петербургская школа физико-математических и компьютерных наук

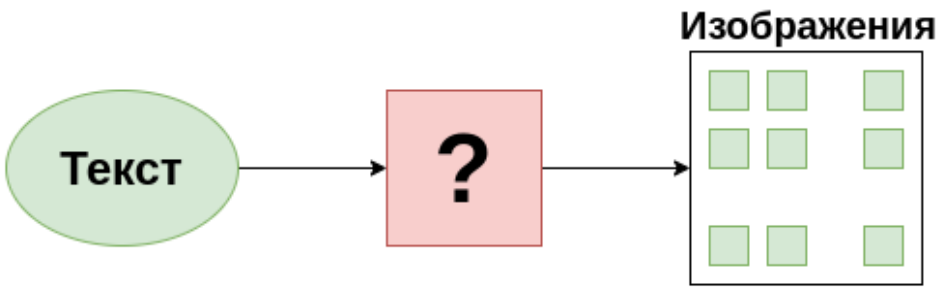
НИУ ВШЭ – Санкт-Петербург

Вторник, 9 июня 2020 года

Введение: Кросс-модальный поиск

Кросс-модальный поиск:

- Поиск **изображений** по запросу-**тексту**
- Поиск **текста** по запросу-**изображению**



Обзор литературы

- **VSE++**[1] — построение векторных представлений объектов.
- **Unsupervised cross-modal retrieval through adversarial learning**[2] — выравнивание векторных представлений.
- **jWAE**[3] — реконструкция объектов из векторных представлений.

---

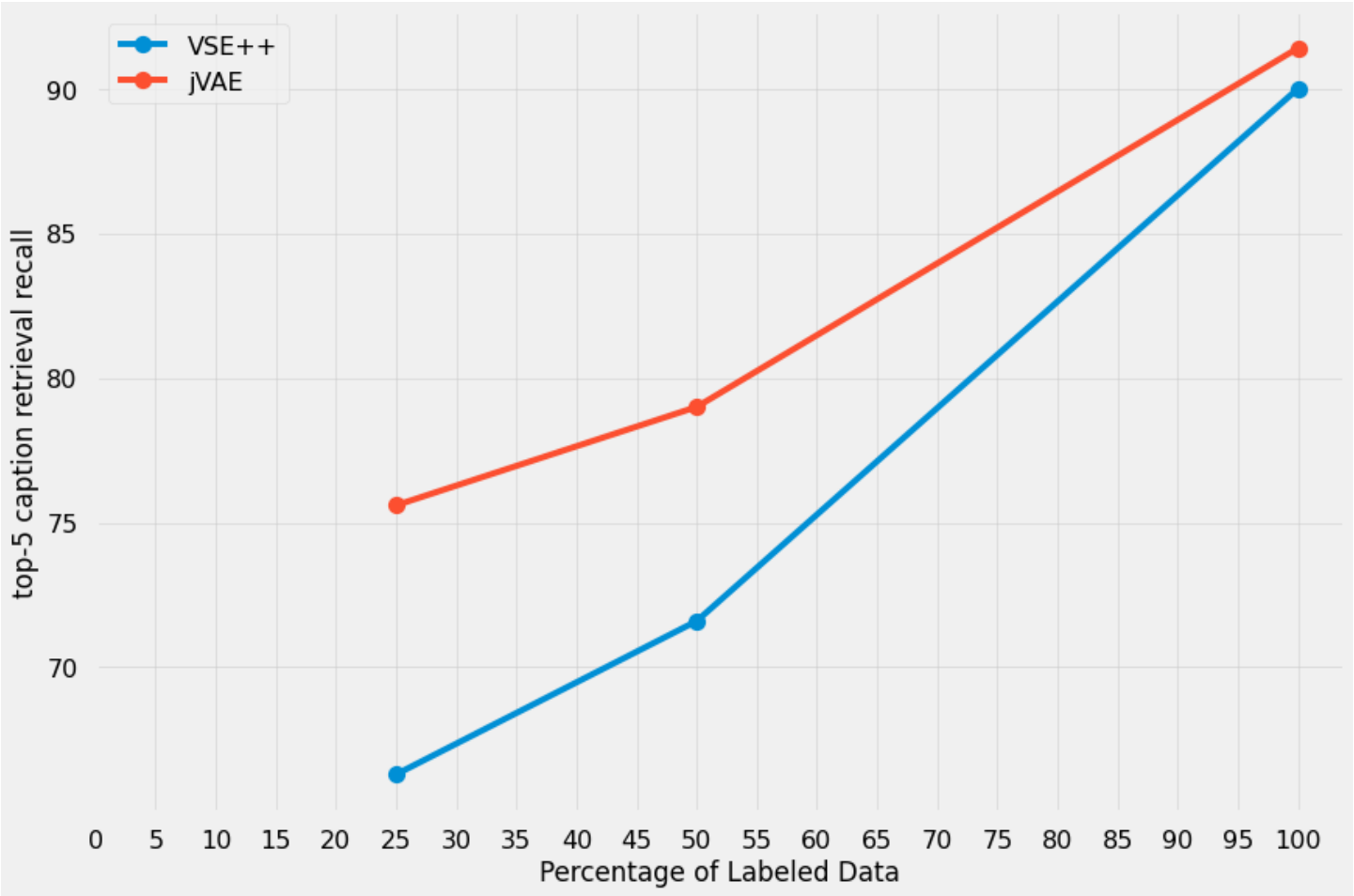
[1] Faghri (2017). VSE++: Improved Visual-Semantic Embeddings.

[2] He, Li Xu (2017). Unsupervised cross-modal retrieval through adversarial learning.

[3] Mahajan (2019). Joint Wasserstein Autoencoders for Aligning Multimodal Embeddings.

Обзор: Проблема решений

Существующие решения требуют большего объема парных данных. Парные данные — дорогие.



Цели и задачи

Цель:

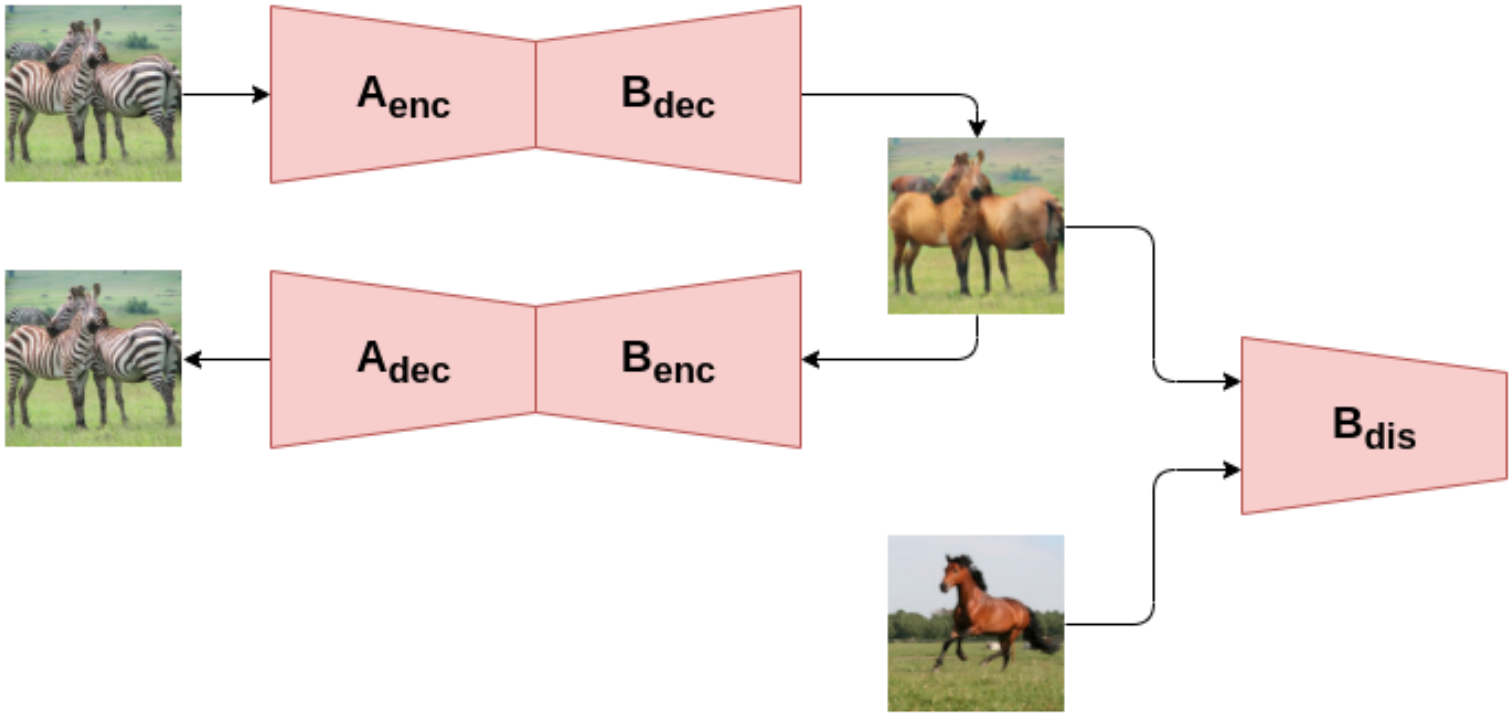
- Построение модели для задачи кросс-модального поиска с меньшей чувствительностью к уменьшению объема парных данных

Задачи:

- Сконструировать архитектуру модели
- Разработать сценарии обучения
- Обучить модель
- Сравнить результаты с существующими аналогами

Задача: Архитектура

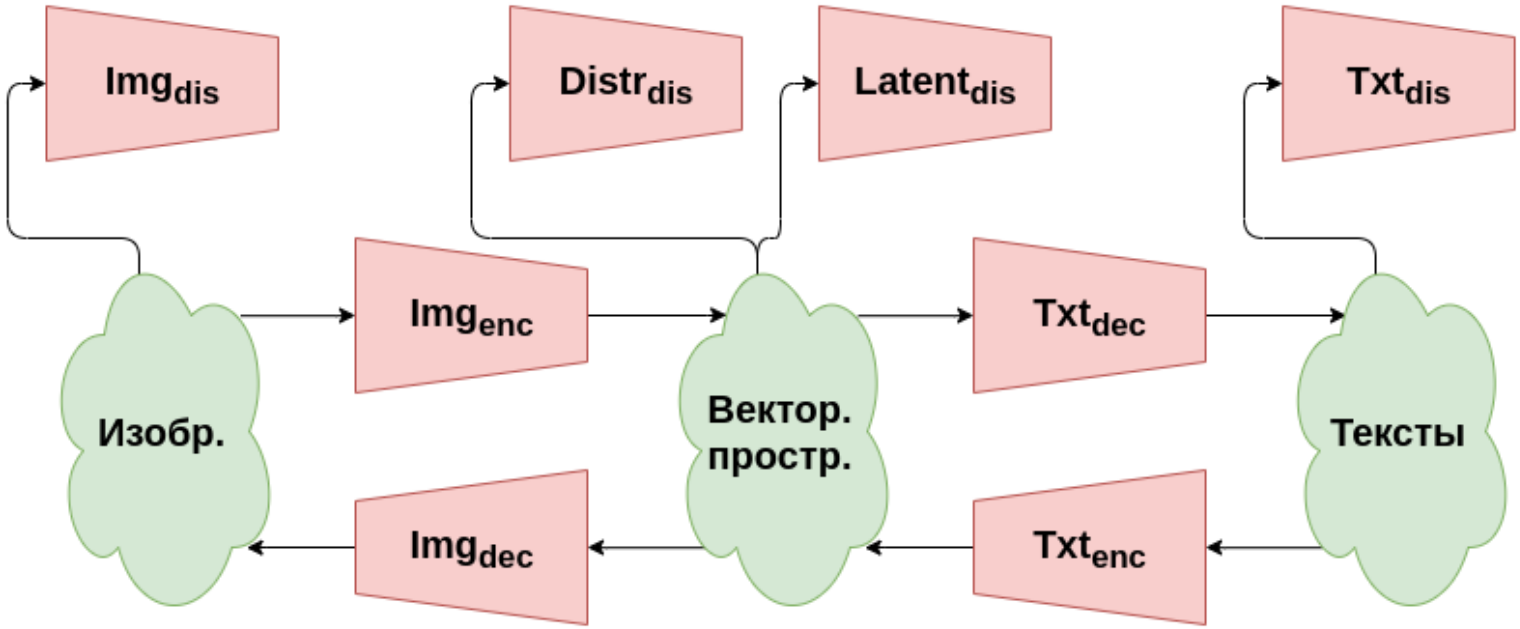
Cycle-Consistency — CycleGAN[4]:



[4]Zhu, Jun-Yan Park, Taesung Isola, Phillip Efros, Alexei. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2242-2251. 10.1109/ICCV.2017.244.

Задача: Сценарии обучения

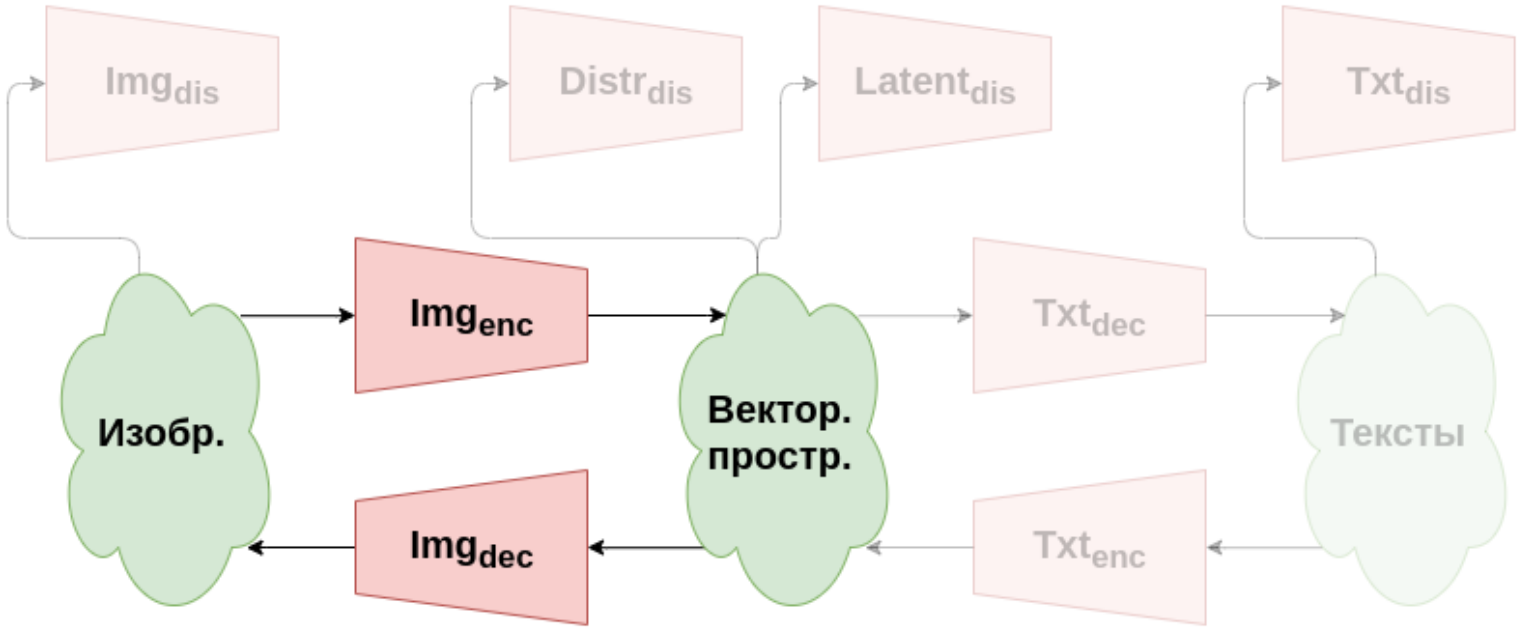
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

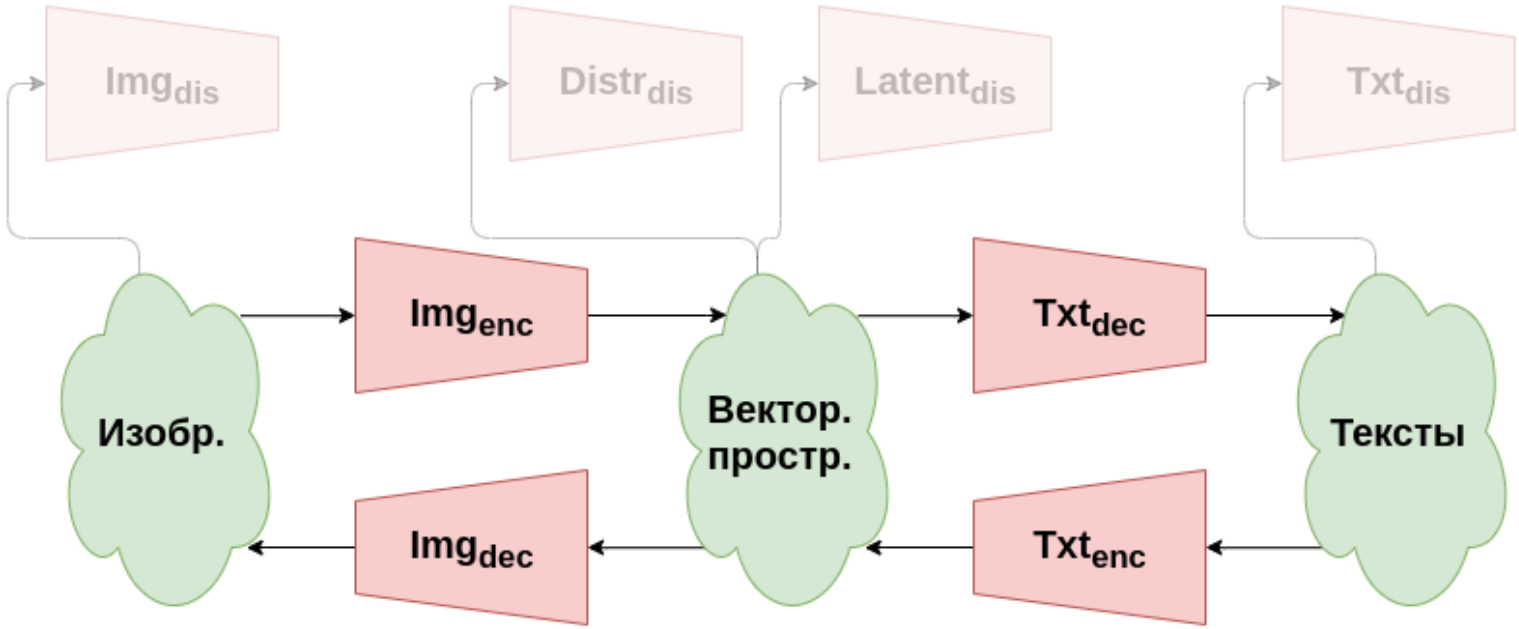
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

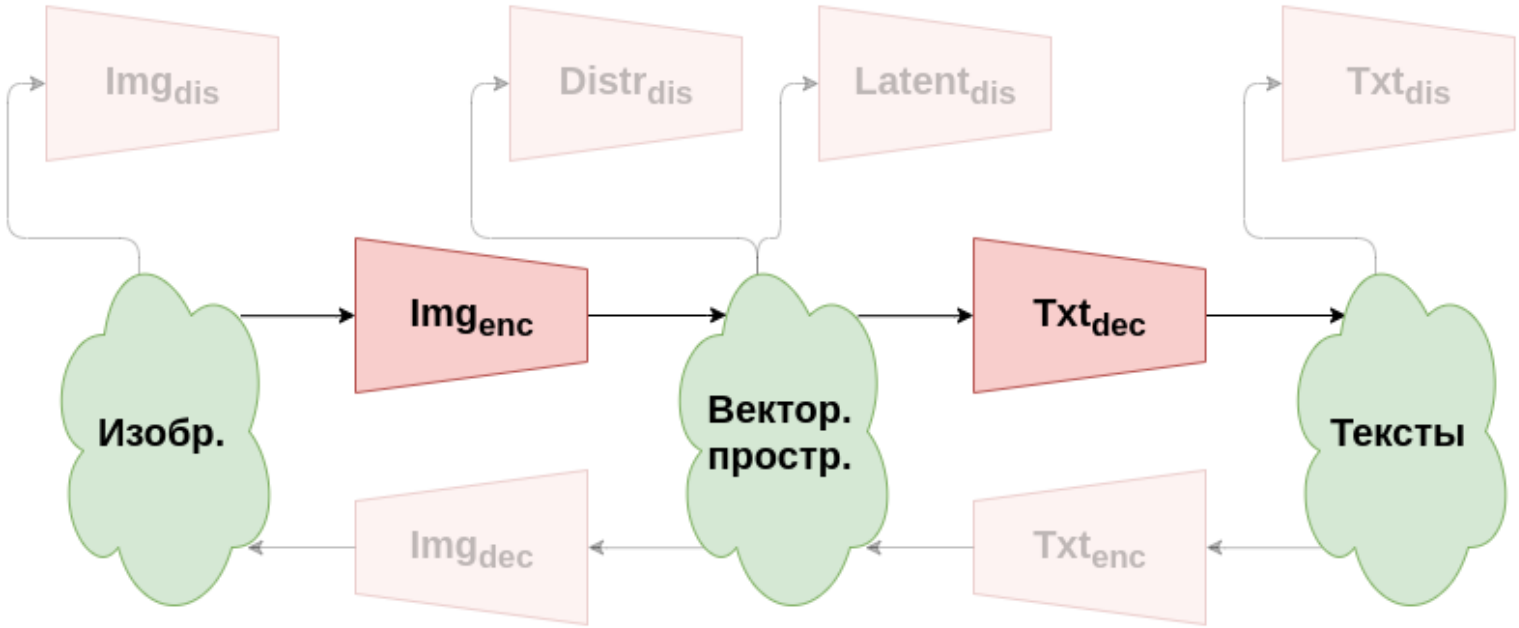
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

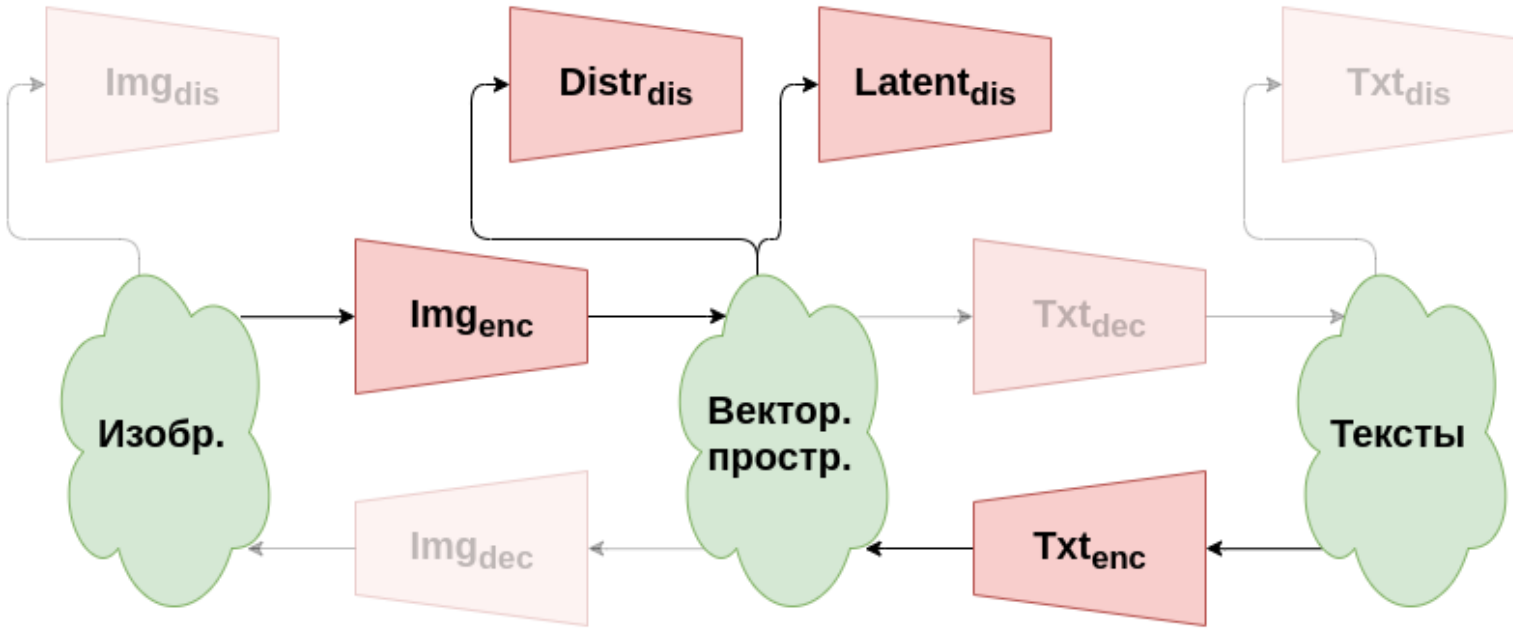
- Автореконструкция
- Циклическая реконструкция
- **Парная реконструкция**
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

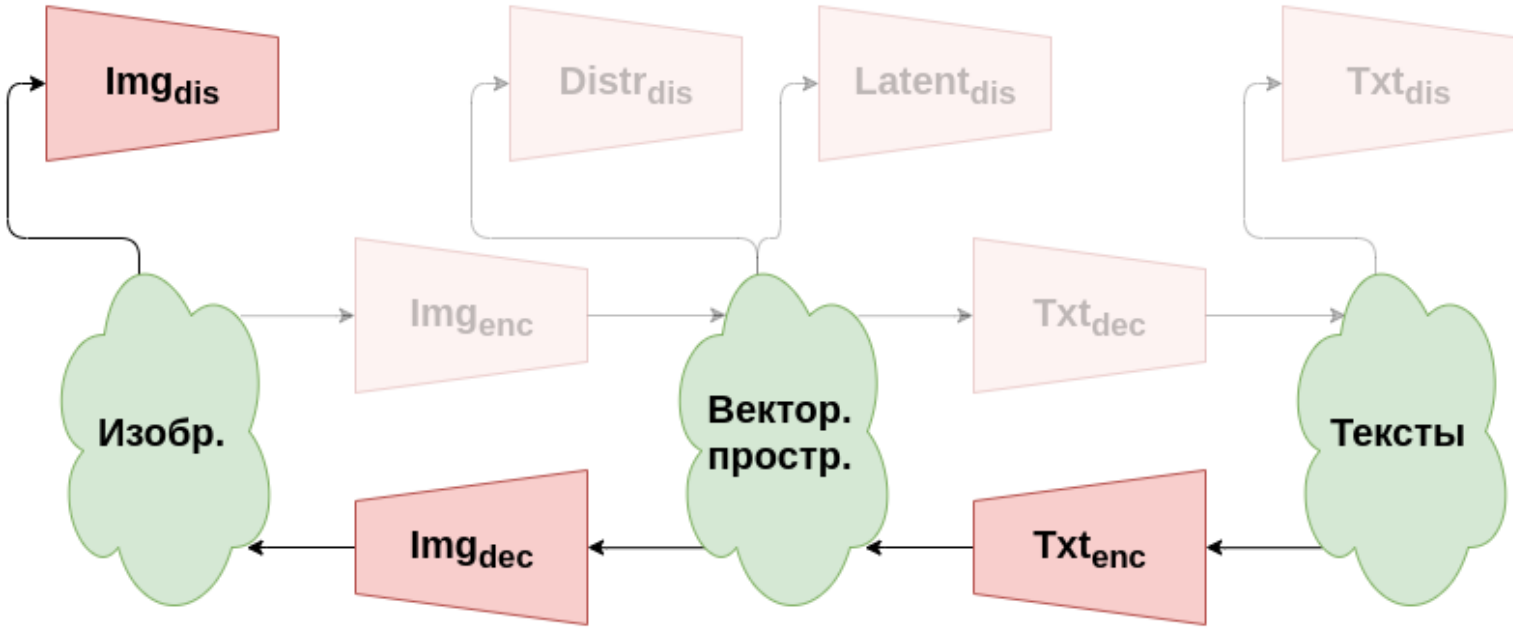
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

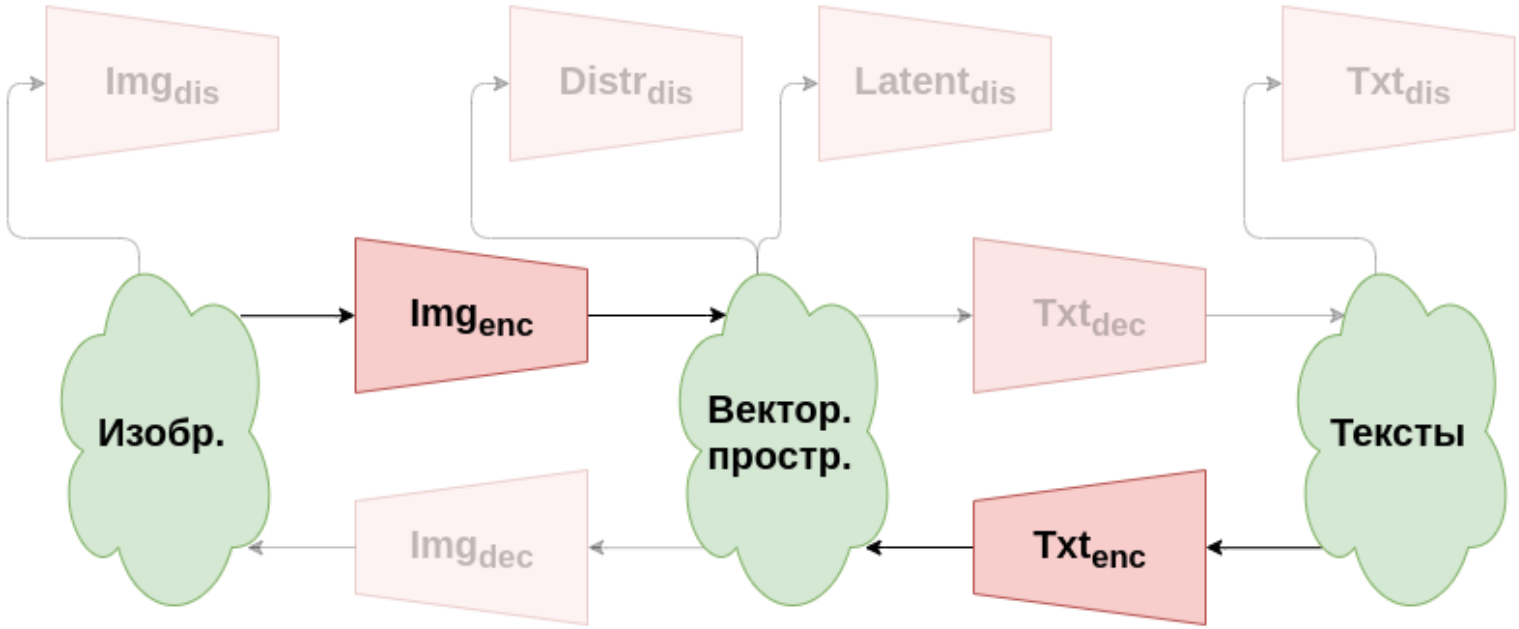
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

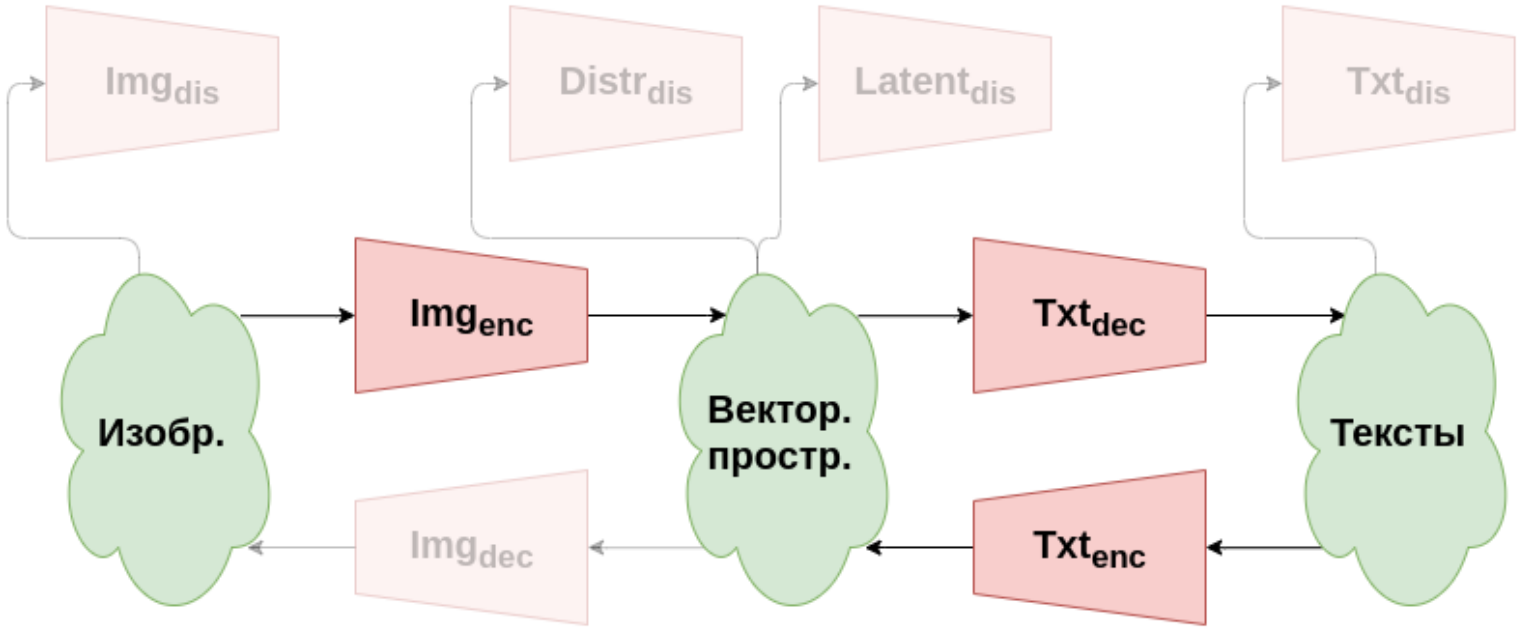
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

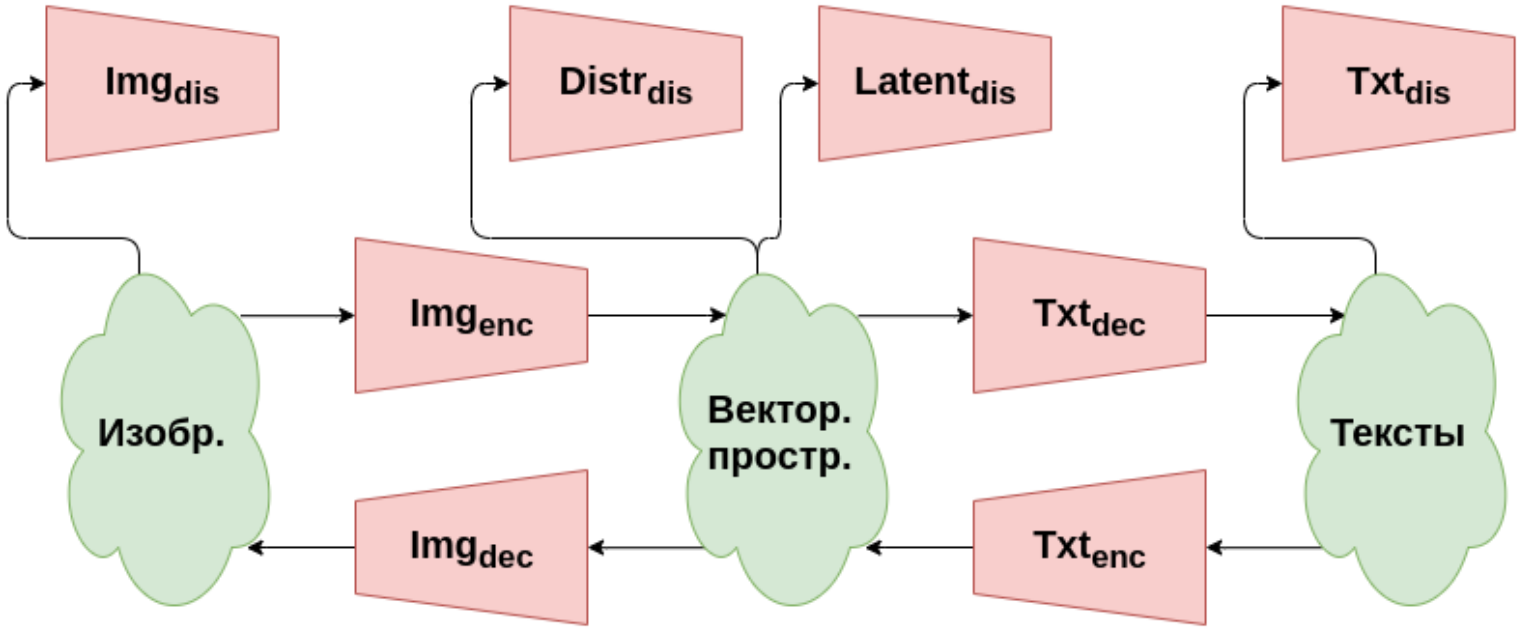
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Сценарии обучения

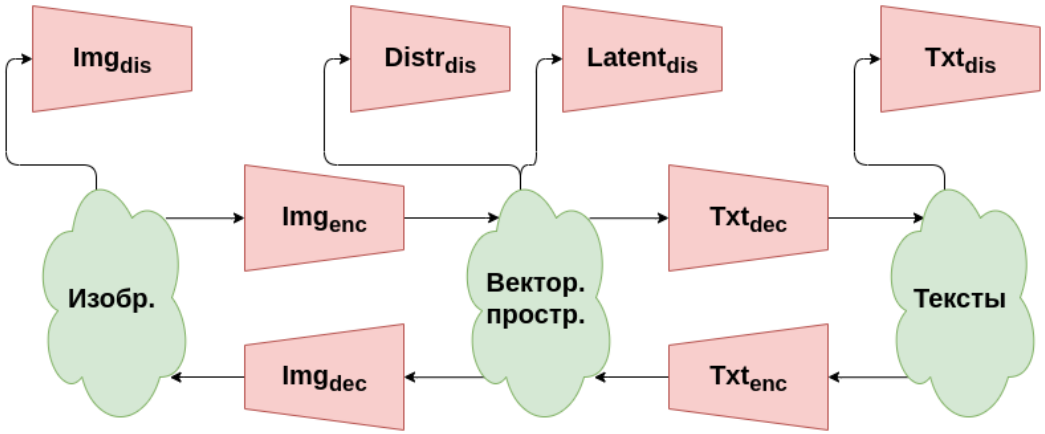
- Автореконструкция
- Циклическая реконструкция
- Парная реконструкция
- Дискриминаторы латентного пространства
- Доменные дискриминаторы
- Парный поиск
- Непарный поиск



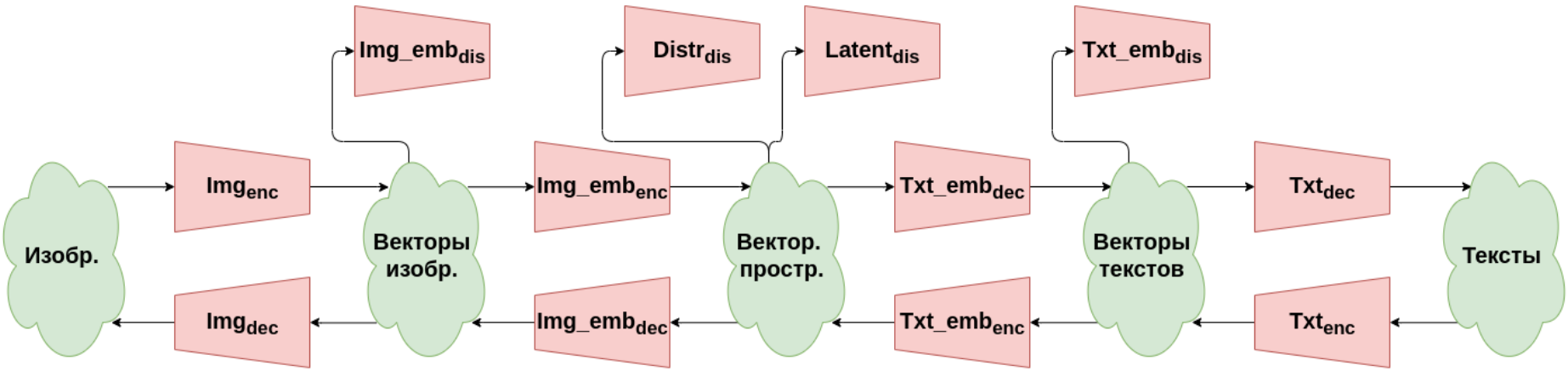
- **Проблема 1:** Низкое качества реконструкции изображений
- **Проблема 2:** Недифференцируемость восстановления текста

Задача: Улучшить архитектуру

Начальная архитектура:

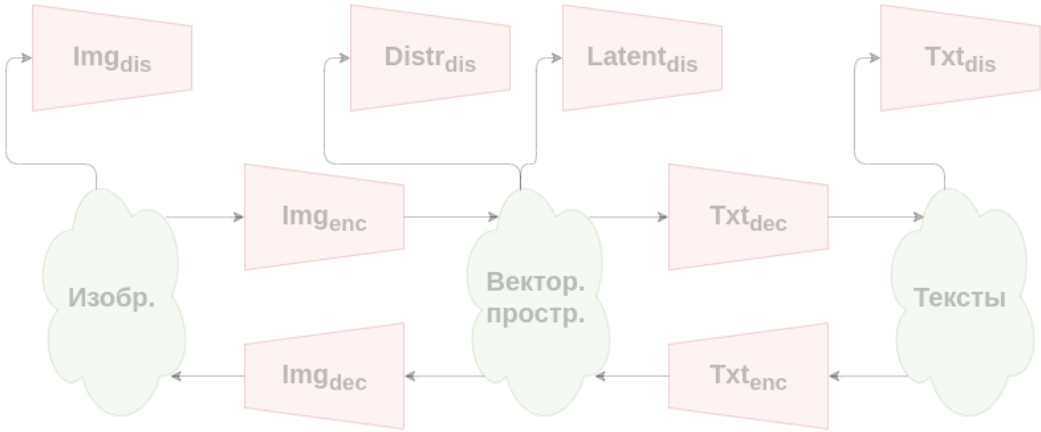


Улучшенная архитектура:

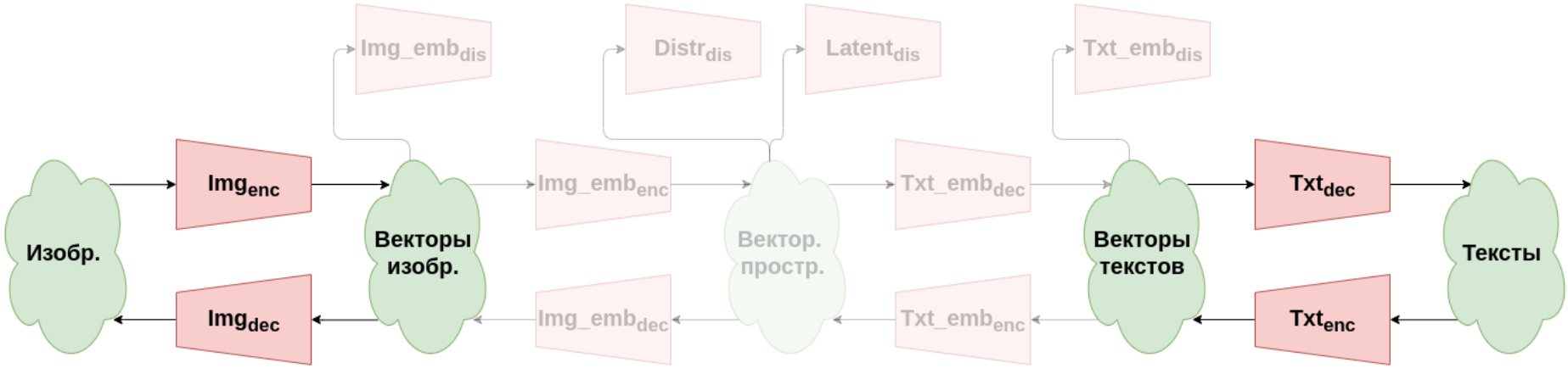


Задача: Улучшить архитектуру

Начальная архитектура:

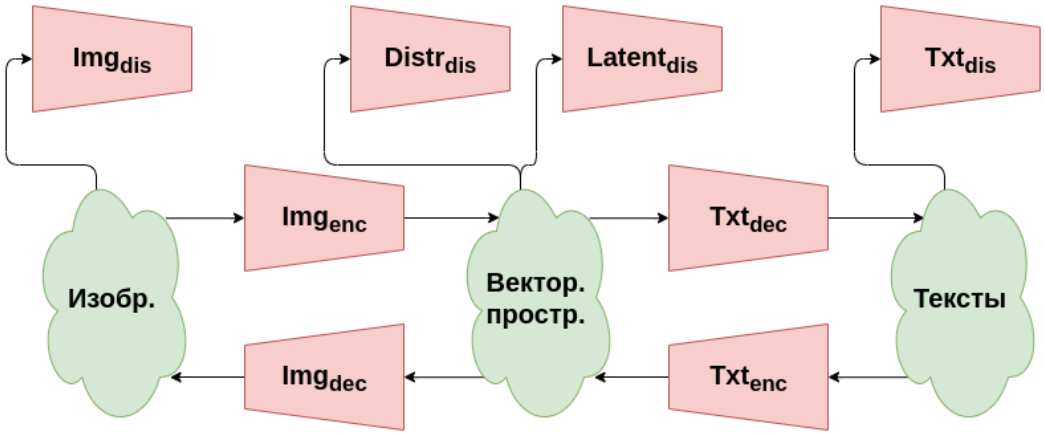


Улучшенная архитектура:

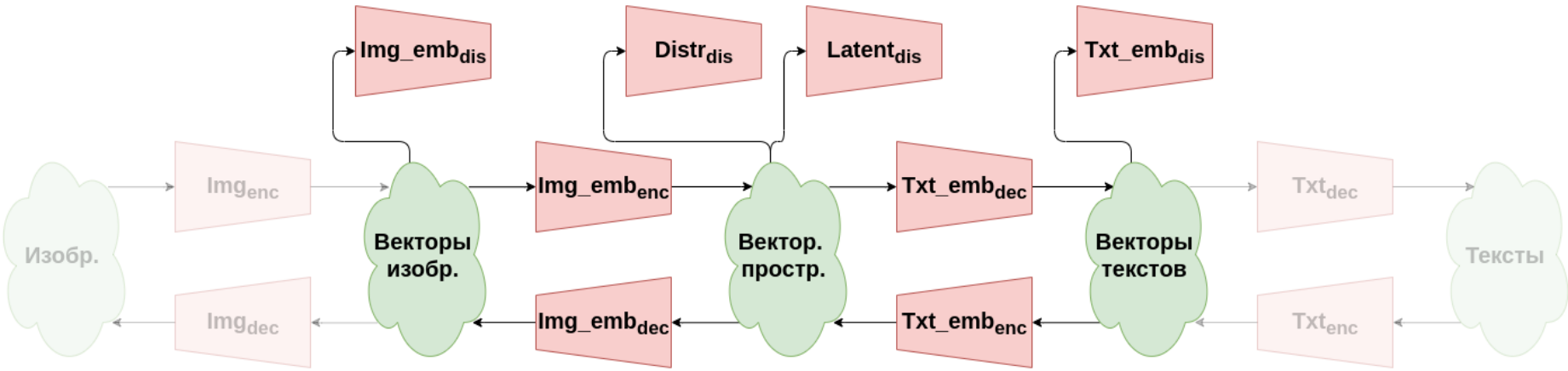


Задача: Улучшить архитектуру

Начальная архитектура:

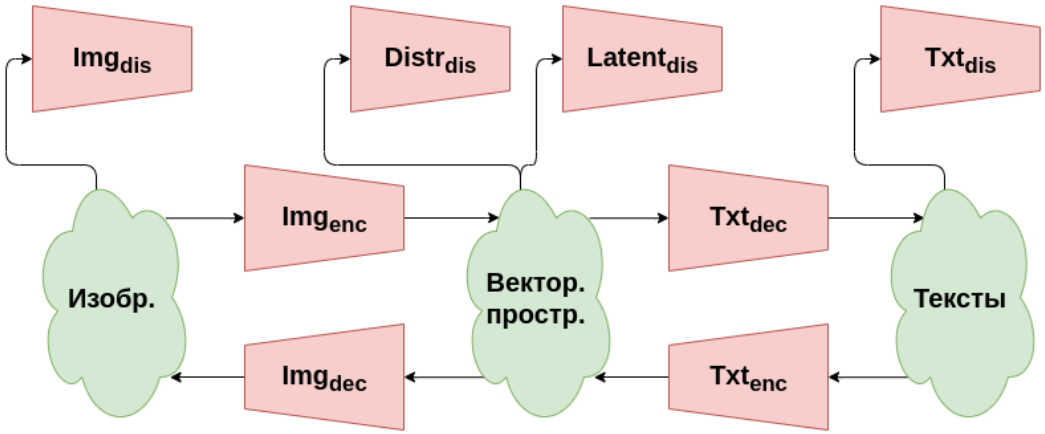


Улучшенная архитектура:

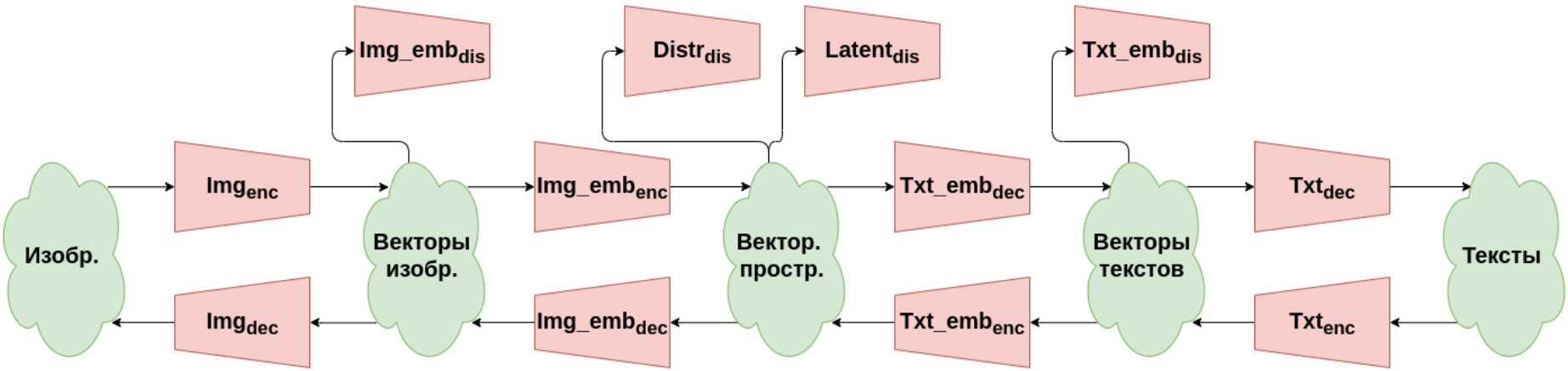


Задача: Улучшить архитектуру

Начальная архитектура:



Улучшенная архитектура:

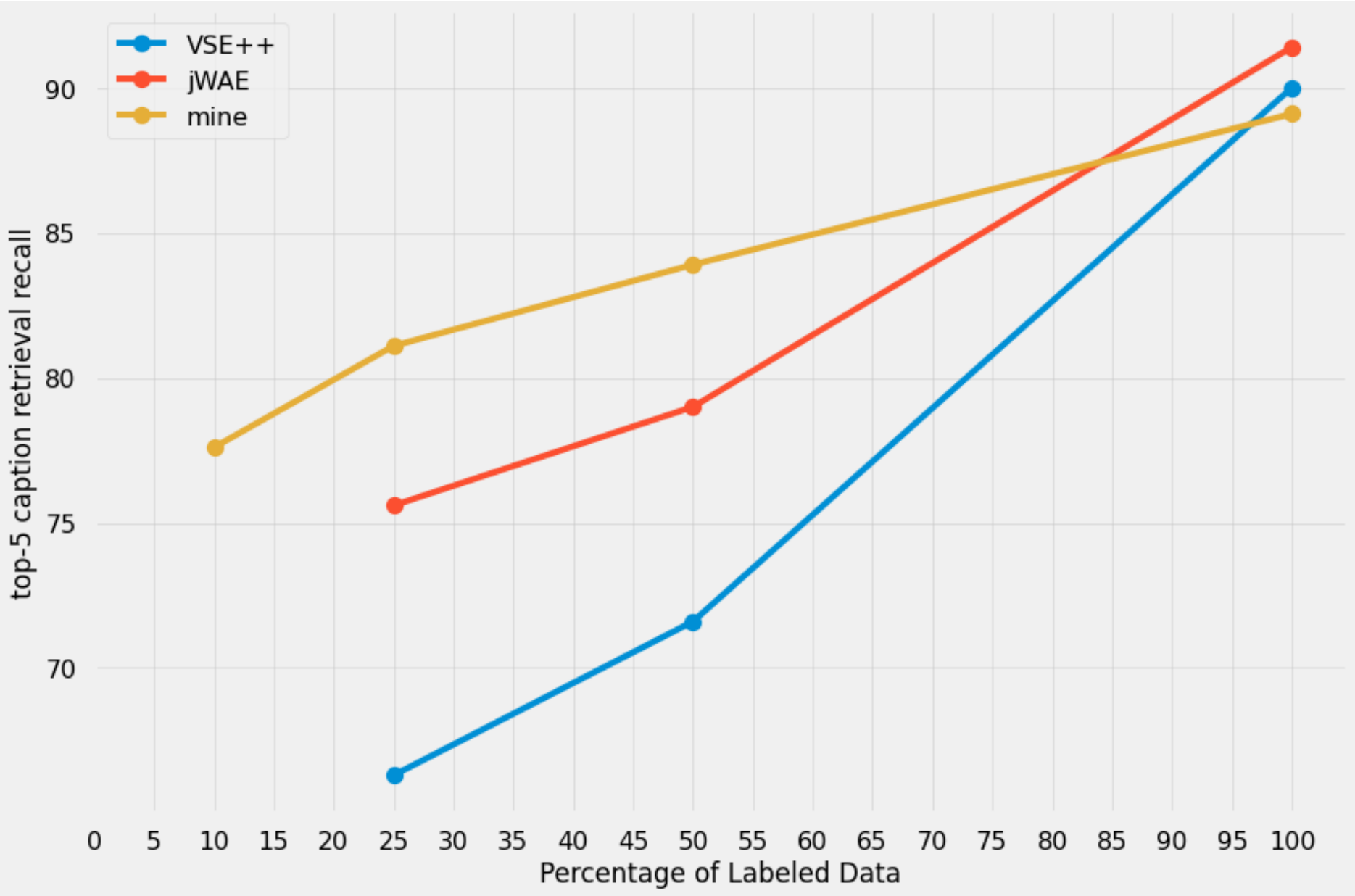


Задача: Обучение

- Датасет: *MS COCO*[5] *2014*, содержит более 80K подписанных изображений.
- Модель состоит из 12 сильно взаимосвязанных частей.
- Обучение происходит поэтапно.
- Для каждого блока было проведено от 10 до 100 экспериментов с архитектурой.

[5] Lin, Tsung-Yi Maire, Michael Belongie, Serge Hays, James Perona, Pietro Ramanan, Deva Dollár, Piotr Zitnick, C.. (2014). Microsoft COCO: Common Objects in Context. 8693. 10.1007/978-3-319-10602-1_48.

Задача: Сравнение — I



Задача: Сравнение — II

Model name	**Parallel part, %**	caption retrieval			image retrieval		
		R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
VSE++[1]	100	64.6	90.0	95.7	52.0	84.3	**92.0**
jWAE[3]	100	**66.6**	**91.4**	**96.6**	**53.1**	**84.5**	**92.0**
mine	100	61.9	89.1	95.1	50.8	82.8	90.2
VSE++[1]	50	-	71.6	-	-	-	-
jWAE[3]	50	-	79.0	-	-	-	-
mine	50	57.5	**83.9**	90.4	47.4	78.4	85.4
VSE++[1]	25	-	66.3	-	-	-	-
jWAE[3]	25	-	75.6	-	-	-	-
mine	25	55.2	**81.1**	87.9	45.6	76.1	82.8
mine	10	52.2	77.6	84.7	43.3	73.1	79.5

Результаты

- Подход Cycle-Consistency приводит к усилению обобщающей способности модели.
- Модель достигает близких к state-of-the-art результатов на полностью парных данных.
- Модель значительно превосходит аналоги на частично парных данных.
- Пока не удалось восстановить связь между текстовой и визуальной информацией вообще без внешней информации о связи.
- Готовится публикация

github.com/GrigoryBartosh/hse08_unimodel

Ссылки

[1] Faghri, Fartash Fleet, David Kiros, Jamie Fidler, Sanja. (2017). VSE++: Improved Visual-Semantic Embeddings.

[2] He, Li Xu, Xing Lu, Huimin Yang, Yang Shen, Fumin Shen, Heng. (2017). Unsupervised cross-modal retrieval through adversarial learning. 1153-1158. 10.1109/ICME.2017.8019549.

[3] Mahajan, Shweta Botschen, Teresa Gurevych, Iryna Roth, Stefan. (2019). Joint Wasserstein Autoencoders for Aligning Multimodal Embeddings. 4561-4570. 10.1109/ICCVW.2019.00557.

[4] Zhu, Jun-Yan Park, Taesung Isola, Phillip Efros, Alexei. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2242-2251. 10.1109/ICCV.2017.244.

[5] Lin, Tsung-Yi Maire, Michael Belongie, Serge Hays, James Perona, Pietro Ramanan, Deva Dollár, Piotr Zitnick, C.. (2014). Microsoft COCO: Common Objects in Context. 8693. 10.1007/978-3-319-10602-1_48.