Стабилизация обучения слотовых распутанных представлений для реальных данных

Фролов В.А.
frolov.vladimir@phystech.edu

Московский физико-технический институт
Центр когнитивного моделирования МФТИ

Научный руководитель д.ф.-м.н. А.И. Панов

2025 – Научно-исследовательский семинар

Цель: Разработка перспективного метода в задаче стабилизации обучения slot-based моделей на основе исследования различных подходов к реконструкции сцен и обнаружению объектов на ней.

Задачи:

- изучение и запуск базового решения;
- интегририрование интересных подходов в существующую модель;
- анализ результатов, релевантный исследуемой проблеме.

Исследуемая проблема:

1. масштабирование объектно-ориентированных методов на сложные данные реального мира, в том числе видео.

Список литературы

1. *M. Seitzer, M. Horn, A. Zadaianchuk, D. Zietlow, T. Xiao, C.J. Simon-Gabriel, T. He, Z. Zhang, B. Schölkopf, T. Brox, F. Locatello.* Bridging the Gap to Real-World Object-Centric Learning // In ICLR 2023.
2. *D. Kirilenko, V. Vorobyov, A. Kovalev, A. Panov.* Object-Centric Learning with Slot Mixture Module // In ICLR 2024.
3. *C. Kim, W. Han, D. Ju, S.J. Hwang.* EAGLE: Eigen Aggregation Learning for Object-Centric Unsupervised Semantic Segmentation // In CVPR 2024.
4. *G. Hinton, O. Vinyals, J. Dean.* Distilling the Knowledge in a Neural Network // In NIPS 2014 Deep Learning Workshop.

План доклада

1. Общая постановка задачи
2. **DINO** and **S**lot **A**ttention **U**sing **R**eal-world data (DINOSAUR)
3. **S**lot **M**ixture **M**odule (SMM)
4. **R**obust **A**daptive **P**robabilistic **I**nference with **DINO** features (RAPID)
5. **S**egment **A**nything **M**odel with **S**lot **A**ttention (SAMSA)

Общая постановка задачи обучения ООП

Задано:

- множество объектов Ω, где $|\Omega| = T$ и $\omega_t \in \Omega$;
- множество признаков $\mathbb{X} \in \mathbb{R}^{H \times W \times 3}$;
- отображение $\varphi : \Omega \to \mathbb{R}^{H \times W \times 3}$, обозначим $\mathbf{x}_t = \varphi(\omega_t)$ и $\forall \mathbf{x}_t,\ \mathbf{x}_t \subset \mathbb{X}$.

Пусть на множестве признаков задана функция $f_{\mathbf{enc}}\big(\mathbf{x}\big)$, а на множестве латентных векторов – $f_{\mathbf{dec}}\big(\mathbf{x}^*\big)$:

$$f_{\mathbf{enc}} : \mathbb{X} \to \mathbb{X}^*,$$

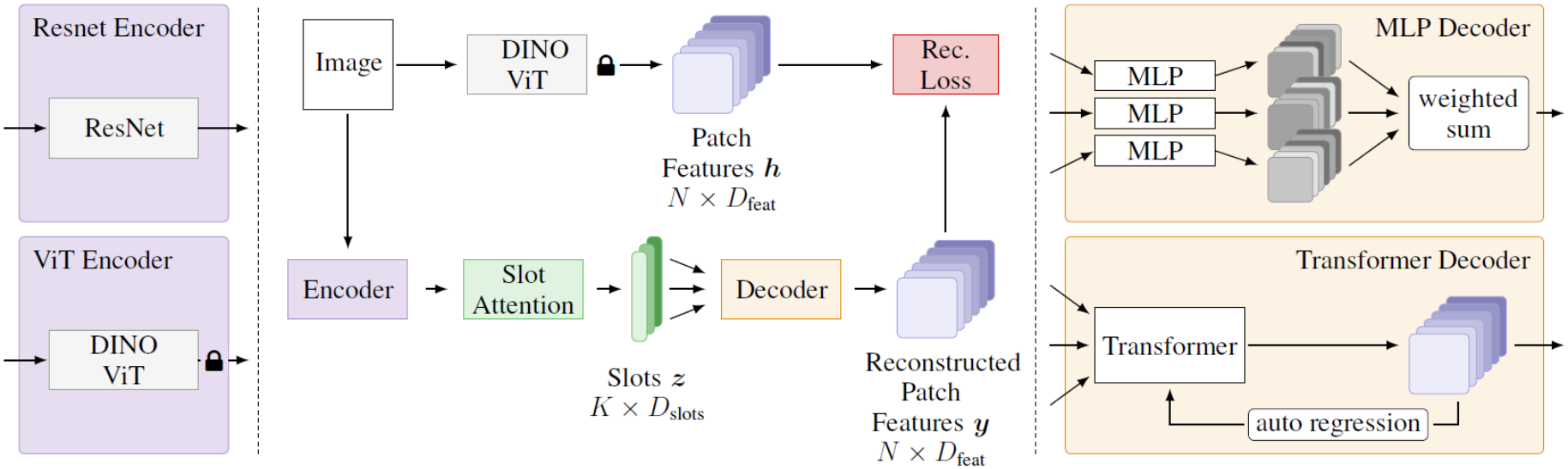
$$f_{\mathbf{dec}} : \mathbb{X}^* \to \hat{\mathbb{X}} \in \mathbb{R}^{H \times W \times 3},$$

где $\mathbb{X}^* \in \mathbb{R}^d$ – представление множества признаков в более компактном латентном пространстве размерности d; $\hat{\mathbb{X}}$ – признаковые описания объектов на выходе, которые мы хотим приблизить к исходным. Для удобства обозначим $f_{\mathbf{dec}}\big(\mathbb{X}^*\big) = \mathcal{P}$.

Заметим, что это задача регрессии, причём такая, что на промежуточном этапе, который будет являться узким местом в модели, может решаться задача классификации или кластеризации для разделения сущностей в объекте.

Оптимизационная задача:

$$\mathcal{Q} = \min_{\mathbf{p}_t \in \mathcal{P}} \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{p}_t\|_2^2.$$



Для каждого изображения $\mathbf{x}_t$, где $t \in \{1, ..., T\}$, ставится задача кодирования и реконструкции признаков $\mathbf{f}_t$, для чего строится отображения $f_{\mathbf{enc}}$:

$$\mathbf{f}_t = f_{\mathbf{enc}}(\mathbf{x}_t) \in \mathbb{R}^{N \times D_{\mathbf{f}}},$$

где $\mathbf{f}_t$ – представление признаков изображения $\mathbf{x}_t$; N – количество токенов; $D_{\mathbf{f}}$ – размерность признаков.

---

https://openreview.net/pdf?id=b9tUk-f_aG

Задаётся K слотов, представляющих отдельные объекты или их части. Набор слотовых представлений для $\mathbf{f}_t$ обозначим $\mathcal{S}_t = \{s_t^1, ..., s_t^K\} \in \mathbb{R}^{K \times D_\mathbf{s}}$. Мы хотим итеративно обновлять эти представления с помощью cross-attention:

$$\mathcal{A}_t = \underset{K}{\text{softmax}}\left[\frac{1}{\sqrt{D_\mathbf{s}}} q(\mathcal{S}_t) \cdot k(\mathbf{f}_t)^T\right] \in \mathbb{R}^{K \times N}, \quad \mathcal{U}_t = \frac{1}{\mathcal{W}_t} \sum_{k=1}^{K} \mathcal{A}_{k,t} \cdot v(\mathbf{f}_{k,t}) \in \mathbb{R}^{K \times D_\mathbf{s}},$$

$$\hat{s}_t^k = \text{GRU}(u_t^k, s_t^k) \in \hat{\mathcal{S}}_t, \qquad \hat{s}_t^k \leftarrow \hat{s}_t^k + \text{MLP}[\text{LN}(\hat{s}_t^k)],$$

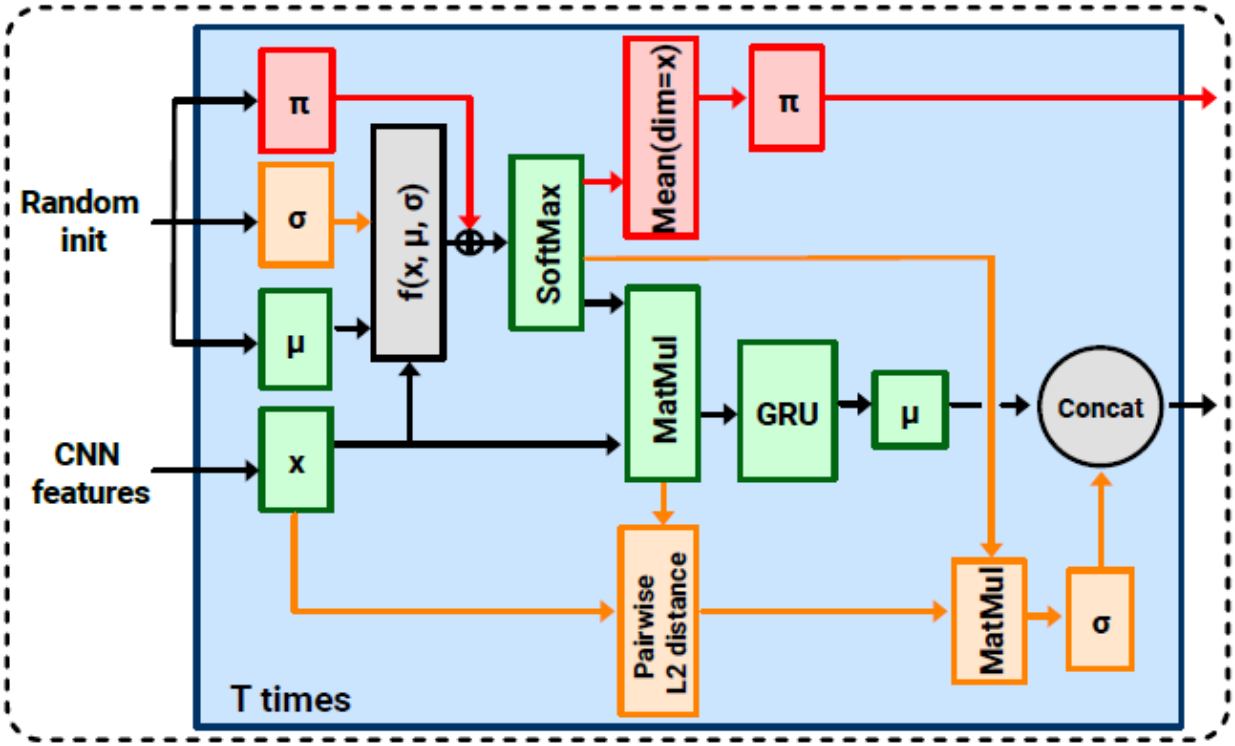
где $\mathcal{W}_t = \sum_{k=1}^{K} \mathcal{A}_{k,t}$; k, q, v – обучаемые линейные проекции $\mathbf{f}_t$, $\mathcal{S}_t$ и $\mathbf{f}_t$, соответственно, которые отображаются в общем пространстве размерности $D_\mathbf{s}$. Обновлённые слотовые представления используются для реконструкции признаков $\mathbf{f}_t$ и предсказания альфа-маски, для чего строится отображение $f_\mathbf{dec}$:

$$\hat{\mathbf{m}}_t^k, \hat{\mathbf{f}}_t^k = f_\mathbf{dec}(\hat{s}_t^k), \quad \hat{\mathbf{m}}_t = \underset{K}{\text{softmax}}(\hat{\mathbf{m}}_t^k), \quad \hat{\mathbf{f}}_t = \sum_{k=1}^{K} \hat{\mathbf{m}}_t \odot \hat{\mathbf{f}}_t^k.$$

Оптимизационная задача:

$$\mathcal{L}_\text{rec} = \min_{\hat{\mathbf{f}}_t \in \mathcal{F}} \sum_{t=1}^{T} \|\mathbf{f}_t - \hat{\mathbf{f}}_t\|_2^2.$$

Slot Mixture Module



Хотим построить вероятностную модель для экземпляра $\mathbf{x}_t$, порождённого смесью распределений с K компонентами и априорными весами $\pi \in \mathbb{R}^K$:

$$\mathbf{x}_t \sim p(\mathbf{x}_t|\theta) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_t|\theta_k), \quad P(\mathbb{X}|\theta) = \prod_{t=1}^{T} p(\mathbf{x}_t|\theta), \quad \sum_{k=1}^{K} \pi_k = 1,$$

https://openreview.net/pdf?id=aBUidW4Nkd

С помощью EM-алгоритма ищется K групп параметров θ_k, задающих каждый экземпляр $\mathbf{x}_t$, которые максимизируют оценку правдоподобия модели $P(\mathbb{X}|\theta)$.

E-шаг:

$$P(\mathbb{X}, \mathbb{X}^*|\theta^*) = \prod_{t=1}^{T}\prod_{k=1}^{K}[\pi_k p(\mathbf{x}_t|\theta_k^*)]^{I(\mathbf{x}_t^*=\mathbf{x}_k^*)},$$

$$Q(\theta^*, \pi^*|\theta, \pi) = \mathbb{E}_{P(\mathbb{X}^*|\mathbb{X},\theta)}[\log P(\mathbb{X}, \mathbb{X}^*|\theta^*)],$$

$$p(\mathbf{x}_k^*|\mathbf{x}_t) = \frac{p(\mathbf{x}_k^*)p(\mathbf{x}_t|\theta_k)}{\sum_{k=1}^{K} p(\mathbf{x}_k^*)p(\mathbf{x}_t|\theta_k)} = \frac{\pi_k \mathcal{N}(\mathbf{x}_t|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_t|\mu_k, \Sigma_k)} = \gamma_{k,t},$$

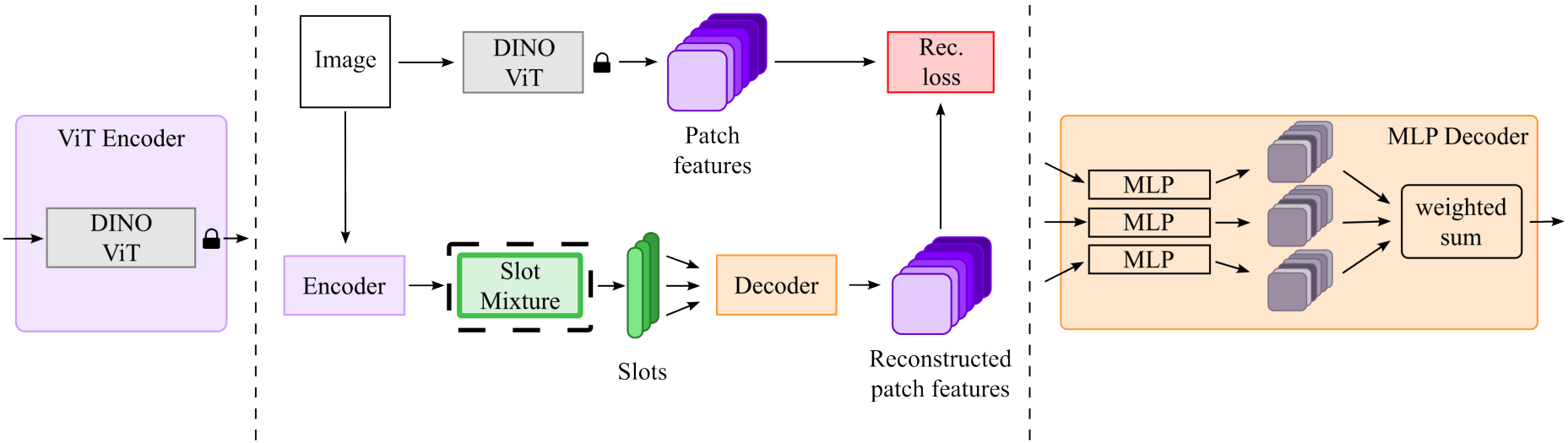
где I – индикаторная функция.

M-шаг:

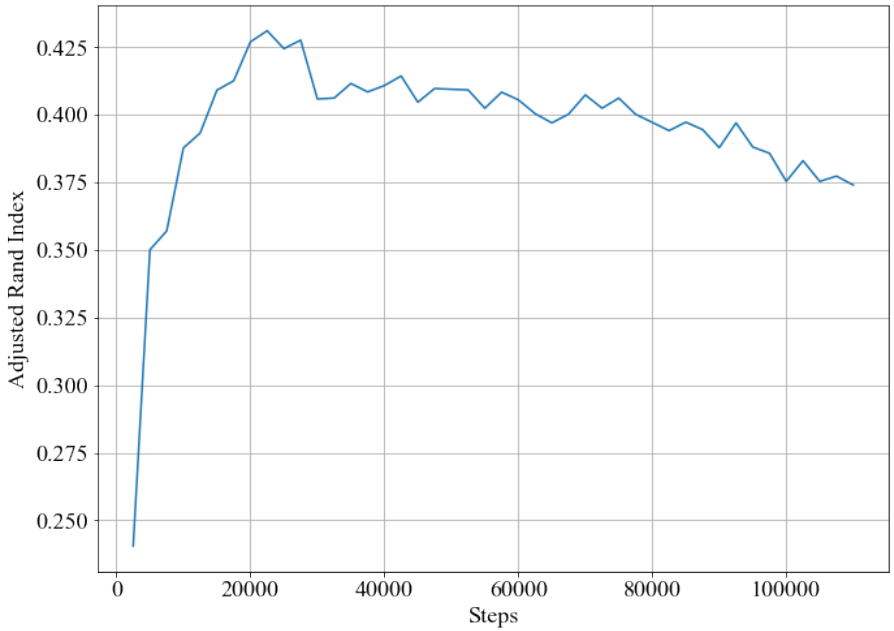
$$(\theta, \pi) = \arg\max_{(\theta^*, \pi^*)} Q(\theta^*, \pi^*|\theta, \pi),$$

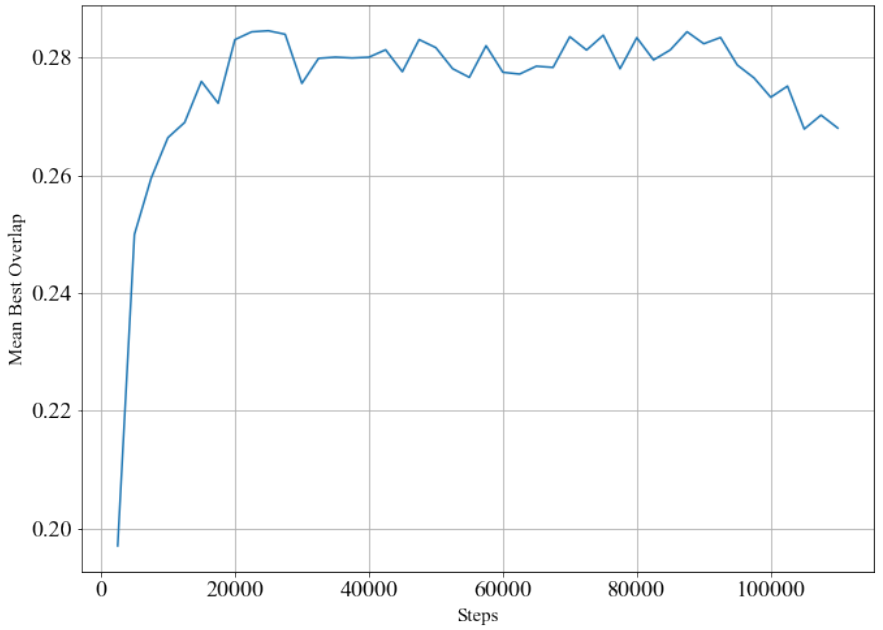
$$\pi_k^* = \frac{\sum_{t=1}^{T}\gamma_{k,t}}{T},\ \mu_k^* = \frac{\sum_{t=1}^{T}\gamma_{k,t}\mathbf{x}_t}{\sum_{t=1}^{T}\gamma_{k,t}},\ \Sigma_k^* = \frac{\sum_{t=1}^{T}\gamma_{k,t}(\mathbf{x}_t - \mu_k)(\mathbf{x}_t - \mu_k)^T}{\sum_{t=1}^{T}\gamma_{k,t}}.$$

Slot Mixture Module



Проблема


Adjusted Rand Index
0.425
0.400
0.375
0.350
0.325
0.300
0.275
0.250
0
20000
40000
60000
80000
100000
Steps


Mean Best Overlap
0.28
0.26
0.24
0.22
0.20
0
20000
40000
60000
80000
100000
Steps

Algorithm 1: Stabilized Slot Mixture Module

Input: $\mathbf{f} \in \mathbb{R}^{N \times D}$ – признаки изображения с добавленными позиционными эмбеддингами;
$\mu, \Sigma \in \mathbb{R}^{K \times D}$, $\pi \in \mathbb{R}^{K}$ – параметры смеси распределений; $\mathcal{I} = \{1, \ldots, L\}$.

Output: $s \in \mathbb{R}^{N \times 2D}$, $\gamma^{**} \in \mathbb{R}^{K \times D}$.

for *each* $i \in \mathcal{I}$ **do**

$\mathcal{D} = f_{\text{PDF}}(\mathbf{f}, \mu, \Sigma, \pi)$;

$\gamma = \frac{\mathcal{D}}{\sum_{k=1}^{K} \mathcal{D}_k}$;

$\mathcal{A} = \frac{\gamma}{\sum_{d=1}^{D} \gamma_d}$;

$\pi^* = \frac{\sum_{t=1}^{T} \mathcal{A}_t}{T}$;

$\mu^* = \frac{\sum_{t=1}^{T} \mathcal{A}_t \mathbf{f}_t}{\sum_{t=1}^{T} \mathcal{A}_t}$;

$\mu^* = \text{GRU}(\text{input} = \mu^*, \text{ hidden} = \mu)$;

$\mu^* = \text{MLP}[\text{LayerNorm}(\mu^*)] + \boxed{\text{LayerNorm}(\mu^*)}$;

$\boxed{\mathcal{D}^* = f_{\text{PDF}}(\mathbf{f}, \mu^*, \Sigma, \pi^*)}$;

$\boxed{\gamma^* = \frac{\mathcal{D}^*}{\sum_{k=1}^{K} \mathcal{D}_k^*}}$;

$\boxed{\mathcal{A}^* = \frac{\gamma^*}{\sum_{d=1}^{D} \gamma_d^*}}$;

$\Sigma^* = \frac{\sum_{t=1}^{T} \mathcal{A}_t^* (\mathbf{f}_t - \mu^*)(\mathbf{f}_t - \mu^*)^T}{\sum_{t=1}^{T} \mathcal{A}_t^*}$;

$\boxed{\pi^{**} = \frac{\sum_{t=1}^{T} \mathcal{A}_t^*}{T}}$;

if $i = L$ **then**

$\boxed{\mathcal{D}^{**} = f_{\text{PDF}}(\mathbf{f}, \mu^*, \Sigma^*, \pi^{**})}$;

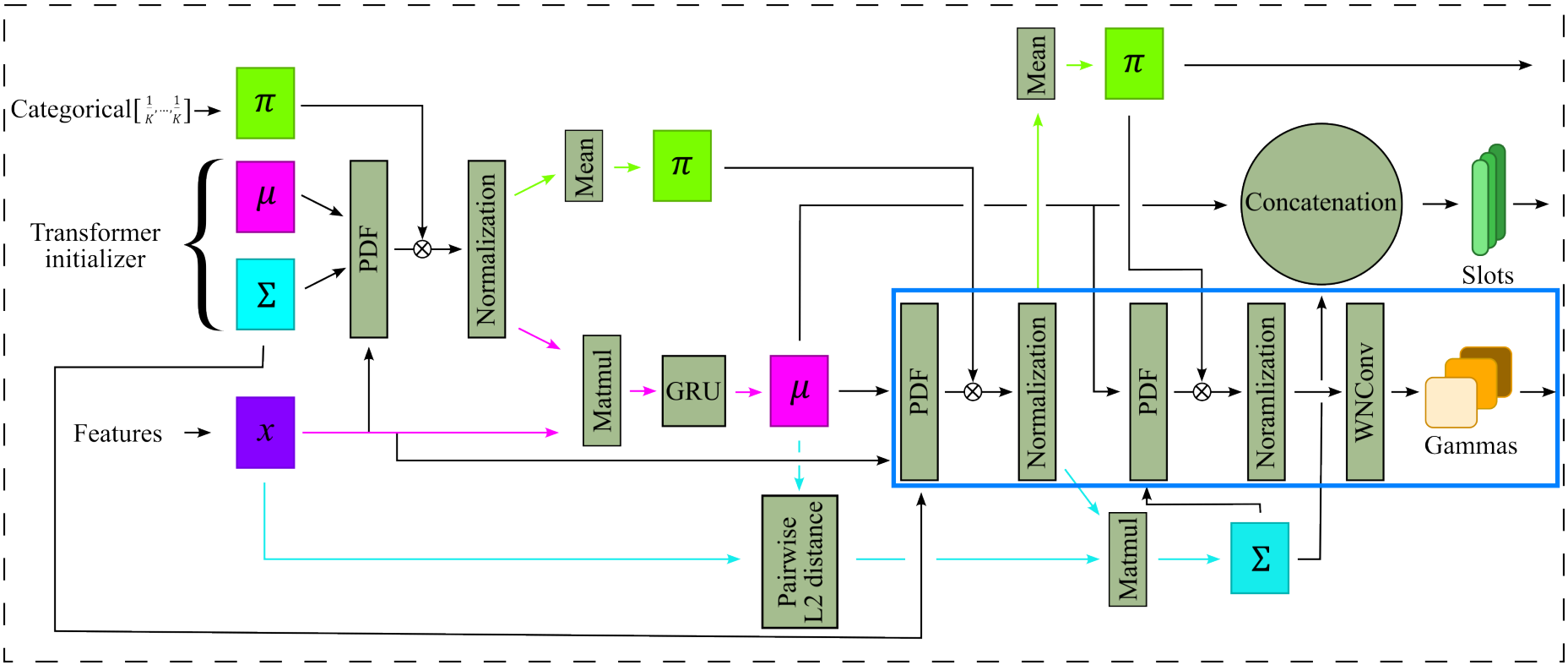
$\boxed{\gamma^{**} = \frac{\mathcal{D}^{**}}{\sum_{k=1}^{K} \mathcal{D}_k^{**}}}$;

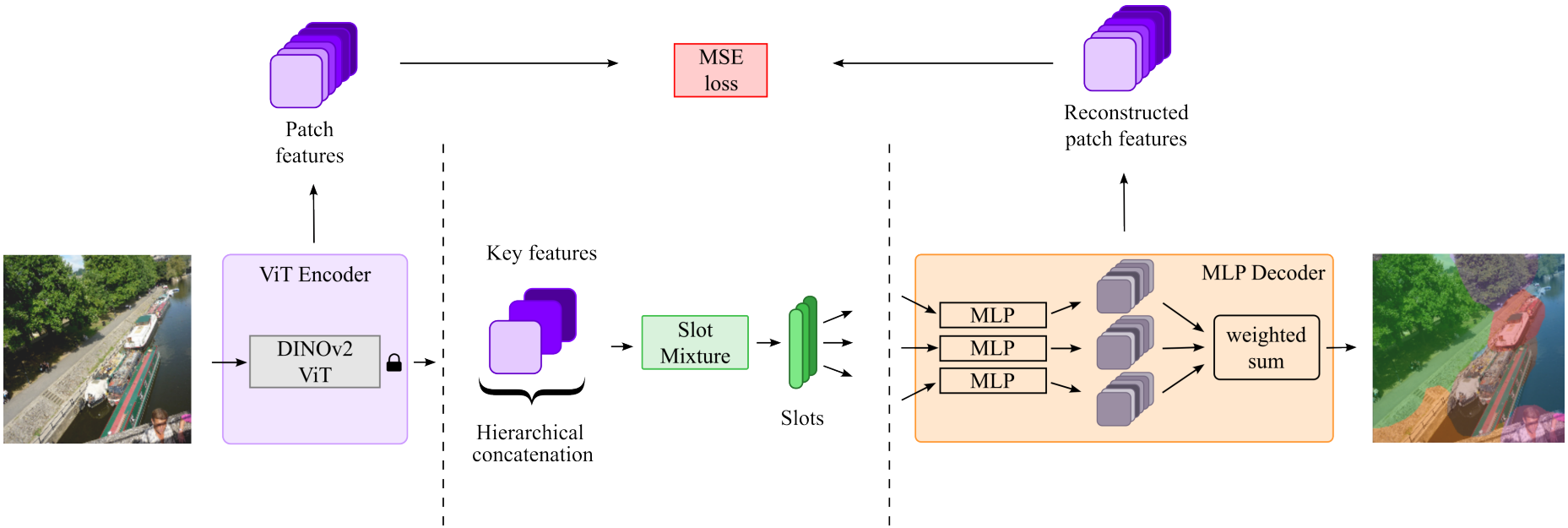
$\boxed{\gamma_{\text{new}} = \text{WNConv}(\gamma^{**})}$;

$s = \text{concat}([\mu^*, \Sigma^*])$;

return s, γ_{new};

Предлагаемый алгоритм




MSE loss
Patch features
Reconstructed patch features
ViT Encoder
DINOv2 ViT
Key features
Hierarchical concatenation
Slot Mixture
Slots
MLP Decoder
MLP
MLP
MLP
weighted sum

Результаты

	COCO		MOVi-C	
	FG-ARI	mBO	FG-ARI	mBO
DINOSAUR	40.5	27.7	68.6	39.1
RAPID	**52.23 ± 0.18**	**30.48 ± 0.09**	**74.7 ± 0.32**	**43.1 ± 0.19**

$$\mathrm{RI} = \frac{2(a+b)}{n(n-1)}, \qquad \mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max \mathrm{RI} - \mathbb{E}[\mathrm{RI}]},$$

где a — число пар объектов, имеющих одинаковые метки в обоих разбиениях (кластеризациях) $\mathbb{Y}$ и $\hat{\mathbb{X}}$; b — число пар объектов, имеющих различные метки в обоих разбиениях; n — число объектов в выборке.

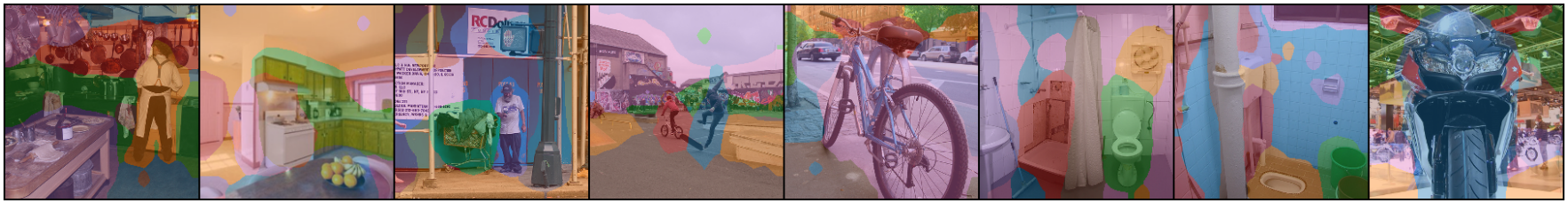
$$\mathrm{mBO} = \frac{1}{TC} \sum_{t=1}^{T} \sum_{c=1}^{C} \frac{\mathbf{y}_{c,t} \cap \arg\max_{\hat{\mathbf{x}}_{c,t} \in \hat{\mathbb{X}}} \mathbf{y}_{c,t} \cap \hat{\mathbf{x}}_{c,t}}{\mathbf{y}_{c,t} \cup \arg\max_{\hat{\mathbf{x}}_{c,t} \in \hat{\mathbb{X}}} \mathbf{y}_{c,t} \cap \hat{\mathbf{x}}_{c,t}},$$

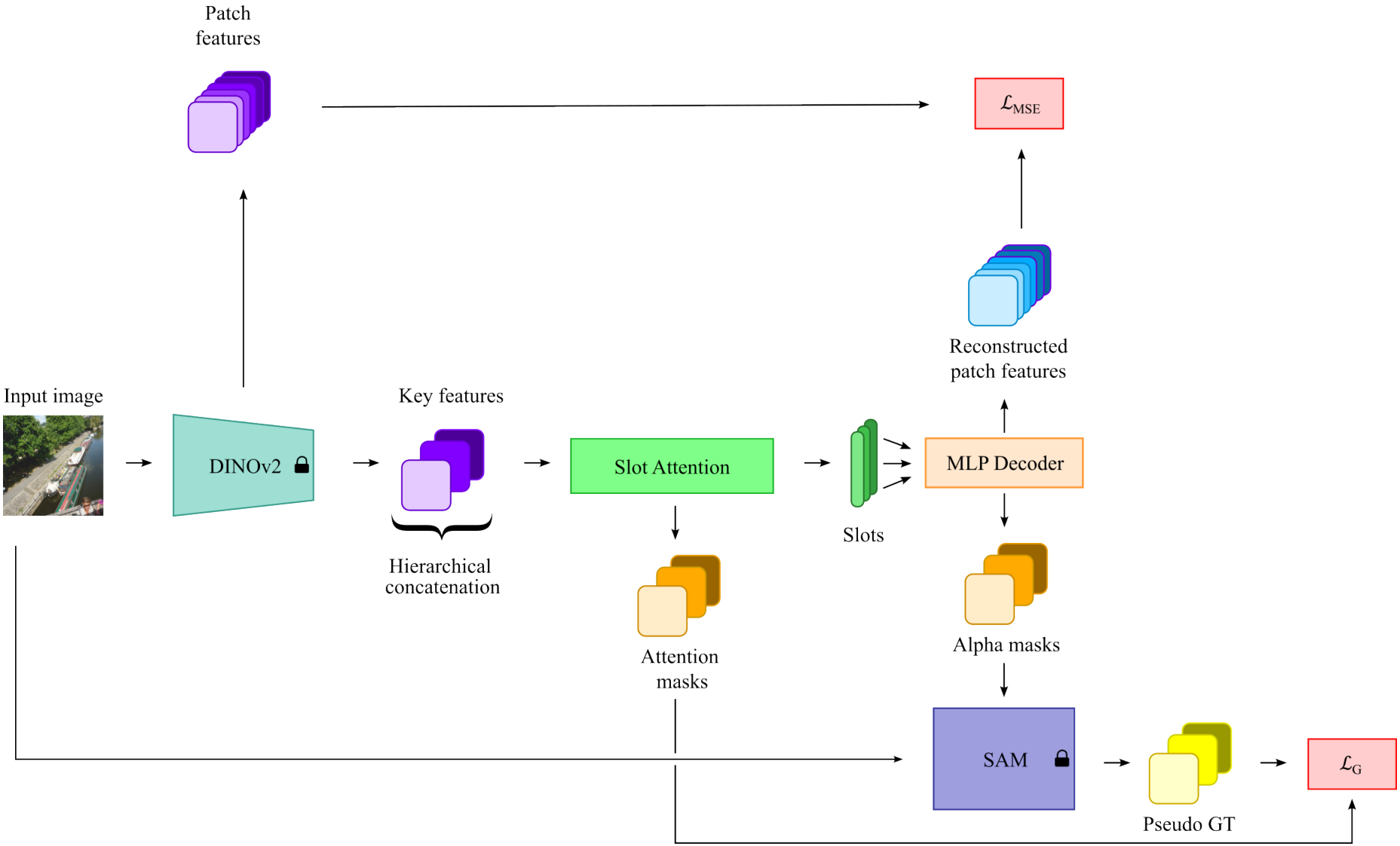
где C — количество классов; $\mathbf{y}_{c,t}$ — истинные метки для экземпляра с порядковым номером t в классе c; $\hat{\mathbf{x}}_{c,t}$ — оценки для экземпляра t в классе c.

DINOSAUR



RAPID





$$\mathcal{L} = \alpha \mathcal{L}_{MSE} + \beta \mathcal{L}_{G}.$$

Algorithm 1: Reference Point Extraction

Input: $\mathbf{m} \in \mathbb{R}^{H \times W}$ – карта коэффициентов внимания, полученная с декодера; n – количество набираемых точек; k – значение перцентиля для вычисления нижнего порога; s_l, s_u – коэффициенты масштабирования стандартного отклонения для нижней и верхней границ.

Output: $p \in \mathbb{R}^{n \times 2}$ – координаты (x, y) опорных точек.

$t_l = \mathcal{P}_k(\mathbf{m})$;
$t_u = \max \mathbf{m}$;
$\mathbf{m}_t = (\mathbf{m} \geq t_l) \cap (\mathbf{m} \leq t_u)$;
$\mu = \text{mean}(\{\mathbf{m}_t = 1\})$;
$\sigma = \text{std}(\{\mathbf{m}_t = 1\})$;
$b_l = \mu - \sigma s_l$;
$b_u = \mu + \sigma s_u$;
$\mathbf{m}_b = (\mathbf{m} \geq b_l) \cap (\mathbf{m} \leq b_u)$;
$p_b = \{(x, y) : \mathbf{m}_b = 1\}$;
$p = \text{random choice}(p_b, n)$;
return p;

Результаты

	FG-ARI	mBO
bboxes	62.28	32.73
points	63.28	**35.19**
points + bboxes	**64.04**	32.40

Original image



Decoder mask with bbox and reference points



Improved mask



Заключение

Планируется:

1. Закончить эксперименты с SAMSA.
2. В качестве возможного решения проблемы выделения отдельных частей объектов видится применение идей из Adaptive SA.
3. Решение проблем с консистентностью слотовых представлений в RAPID.

Спасибо за внимание!

frolov.vladimir@phystech.edu