Гибридный подход к распознаванию рукописного текста

Войт Руслан Александрович д.т.н., проф. Местецкий Л.М.

МГУ имени М.В. Ломоносова, факультет ВМК, кафедра ММП

20 февраля 2026 г.

Задача распознавания рукописного текста

- Построить отображение изображения в последовательность символов:

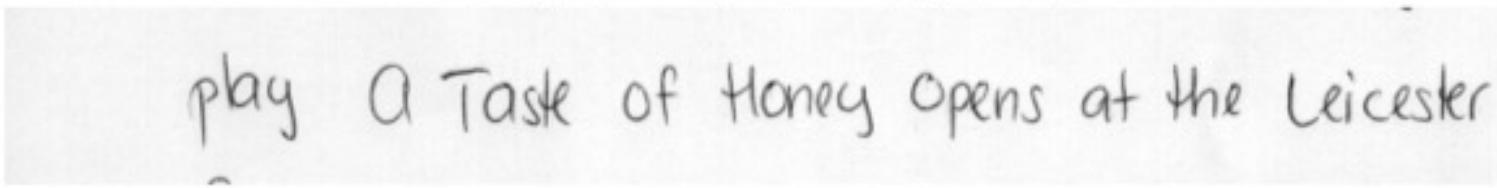
$$I \in \mathbb{R}^{H \times W} \longrightarrow Y = (y_1, \ldots, y_T).$$

- Основные метрики:

$$\text{CER} = \frac{\rho_{\text{Левенштейн}}(Y_{\text{ист}}, Y_{\text{пред}})}{|Y_{\text{ист}}|} \cdot 100\%, \text{WER} = \frac{\rho_{\text{Левенштейн}}(W_{\text{ист}}, W_{\text{пред}})}{\text{число слов}} \cdot 100\%$$

- Сложности: высокая вариативность почерков, шумы, малые объёмы размеченных данных (особенно для исторических документов).

Text: play|A|Taste|of|Honey|opens|at...



Пример из IAM.

Существующие подходы и их ограничения

Растровые модели (CRNN, TrOCR, VAN)

– Достигают высокого качества на больших корпусах (IAM).
– Требуют много размеченных данных, склонны к переобучению на малых выборках.
– Не используют геометрию письма.

Гибридные и структурные подходы (InkSight, GCM, GNN на chain codes)

– Учитывают структуру штрихов, но построение графа часто эвристическое и нестабильное на зашумлённых данных.
– Либо не интегрированы с распознаванием.

Наш вклад

Предлагается двухпоточная архитектура с **математически обоснованным** построением графа штрихов на основе непрерывной морфологии (скелетизация через диаграмму Вороного). Это повышает устойчивость и обобщающую способность.

Математическая постановка задачи

Обучающая выборка: $\mathcal{D} = \{(I^{(n)}, Y^{(n)})\}_{n=1}^{N}$, $I \in \mathbb{R}^{H \times W}$, $Y \in \mathcal{A}^*$. Модель f_θ состоит из:

- визуального энкодера E_v: $V = E_v(I)$,
- графового энкодера E_g: $H = E_g(G)$ (граф G строится по I),
- механизма слияния Φ: $U = \Phi(V, H)$,
- CTC-декодера D: $P(Y|U)$.

Функция потерь CTC:

$$\mathcal{L}_{\text{CTC}}(Y, U) = -\log \sum_{\pi \in \mathcal{B}^{-1}(Y)} \prod_{t} p_t(\pi_t), \quad p_t(\pi_t) = \text{softmax}(W_{\text{ctc}}\, u_t).$$

Оптимизация: $\theta^* = \arg\min_\theta \sum_n \mathcal{L}_{\text{CTC}}(Y^{(n)}, U^{(n)}) + \lambda\Omega(\theta)$.

Общая архитектура гибридной модели

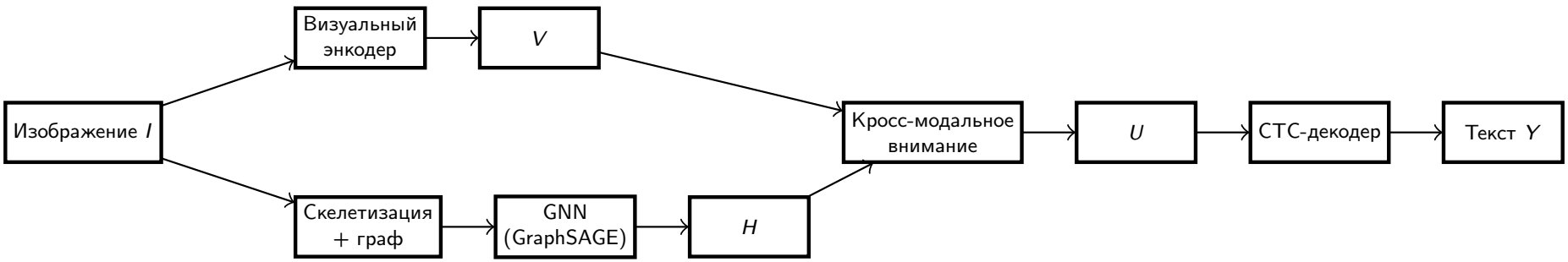


Рис. 1. Двухпоточная архитектура: визуальный поток (VAN/CNN), графовый поток (скелет $\rightarrow$ граф штрихов $\rightarrow$ GNN), кросс-модальное внимание, CTC-декодер.

Визуальный поток

- Используется архитектура **Vertical Attention Network (VAN)** (можно заменить на свёрточную сеть для прототипирования).
- Выход: последовательность визуальных признаков

$$V = (v_1, \ldots, v_M), \quad v_t \in \mathbb{R}^{d_v} \ (d_v = 512).$$

- Внимание по вертикали позволяет фокусироваться на строках и символах.

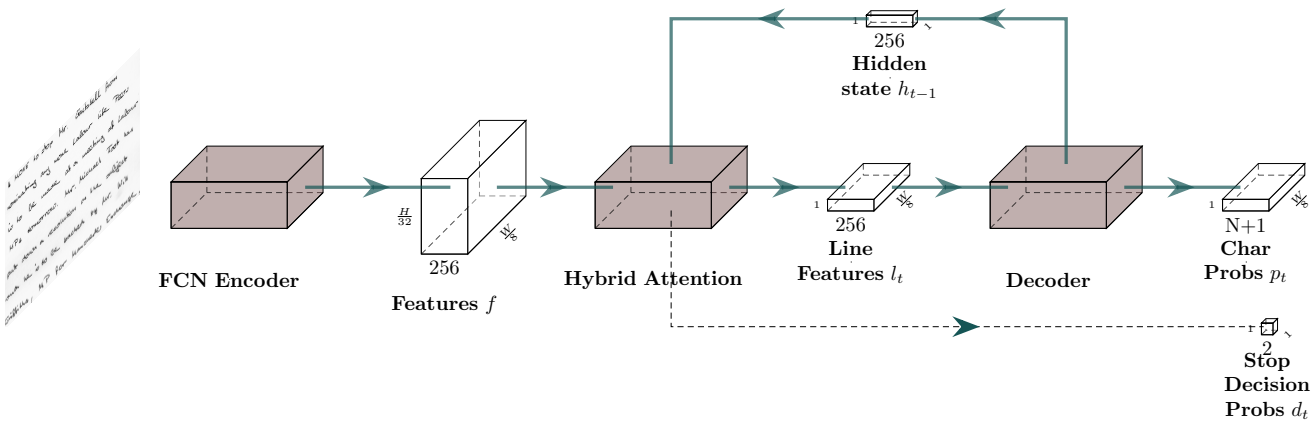


Рис. 2. Схема блока вертикального внимания.

Графовый поток: построение графа штрихов

1. Бинаризация изображения.
2. Построение скелета (медиального представления) методом непрерывной морфологии на основе диаграммы Вороного [1, 2]. Скелет – геометрический граф.
3. Штриховая сегментация: разбиение скелета на элементарные фрагменты (штрихи) по топологическим и геометрическим критериям.
4. Вершины графа G – штрихи; рёбра – общие концевые точки или близость.

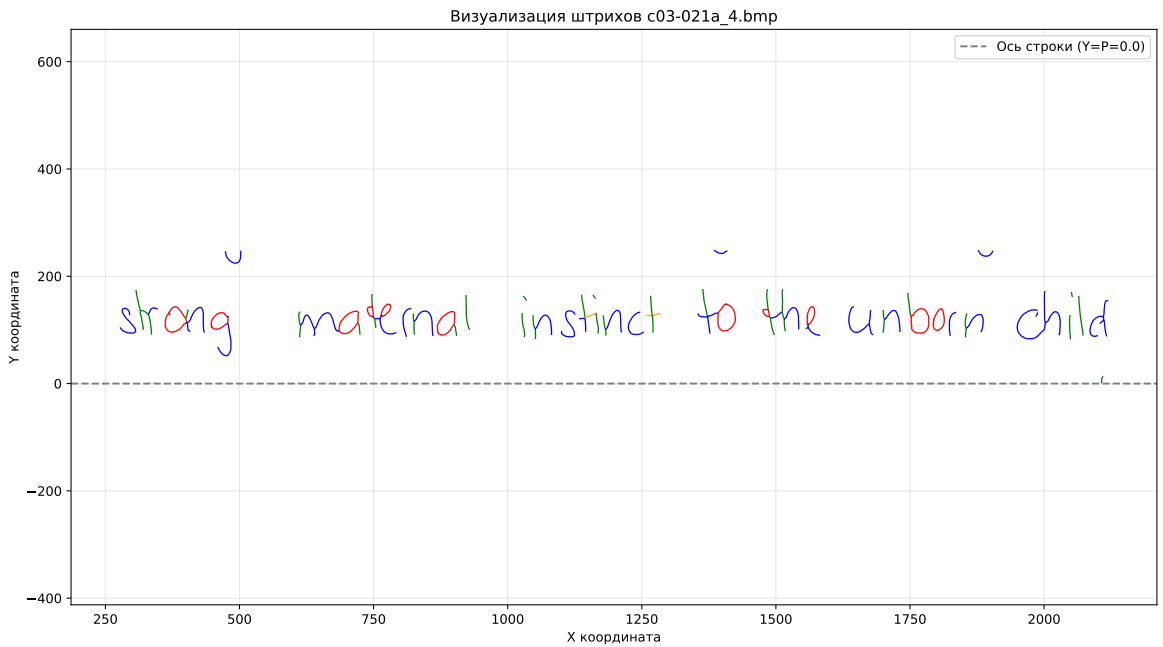


Рис. 3. Пример: исходное слово (слева) и его штриховой граф (справа).

Признаки вершин графа

Каждому штриху s_i сопоставляется вектор признаков:

- геометрические: длина, средняя кривизна, ориентация, координаты центра масс;
- тип штриха (вертикальный/горизонтальный отрезок, дуга, петля) – one-hot;
- параметры из процесса разложения (радиус вписанной окружности для петель).

Итоговая матрица признаков $X \in \mathbb{R}^{K \times d_f}$ ($d_f \sim 20$).

Графовый энкодер (GraphSAGE)

Двухслойная архитектура с агрегацией по соседям:

$$h_i^{(1)} = \mathrm{ReLU}\big(W_1 \cdot \mathrm{CONCAT}\big(f_i,\ \mathrm{MEAN}\{f_j : j \in \mathcal{N}(i)\}\big)\big)\,,$$
$$h_i^{(2)} = \mathrm{ReLU}\Big(W_2 \cdot \mathrm{CONCAT}\big(h_i^{(1)},\ \mathrm{MEAN}\{h_j^{(1)} : j \in \mathcal{N}(i)\}\big)\Big)\,.$$

Выход: $H = (h_1^{(2)}, \ldots, h_K^{(2)})$, $h_i^{(2)} \in \mathbb{R}^{d_h}$ $(d_h = 256)$.

Слияние модальностей (кросс-модальное внимание)

Проекции:

$$Q = VW_Q,\ K = HW_K,\ V_{\text{val}} = HW_V, \quad W_Q \in \mathbb{R}^{d_v \times d_k},\ W_K \in \mathbb{R}^{d_h \times d_k},\ W_V \in \mathbb{R}^{d_h \times d_v}.$$

Матрица внимания:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{M \times K}, \quad \widetilde{H} = AV_{\text{val}} \in \mathbb{R}^{M \times d_v}.$$

Итоговое представление (конкатенация):

$$U = [V \mid \widetilde{H}] \in \mathbb{R}^{M \times 2d_v} \xrightarrow{\text{линейный слой}} U' \in \mathbb{R}^{M \times d_v}.$$

Декодирование и обучение

- CTC-декодер: линейный слой + softmax над алфавитом $\mathcal{A}' = \mathcal{A} \cup \{\epsilon\}$.
- Вероятность последовательности Y:

$$P(Y|U') = \sum_{\pi \in \mathcal{B}^{-1}(Y)} \prod_{t=1}^{M} p_t(\pi_t).$$

- Обучение сквозное, минимизация $\mathcal{L}_{\text{CTC}} + \lambda \|\theta\|_2^2$.
- Оптимизатор Adam, начальный lr $= 10^{-3}$, dropout 0.3.

Эксперименты: настройка

- Датасет: подмножество IAM (90% train, 10% test).
- Конфигурации:
  1. Базовая (только визуальный CNN-энкодер) — *CNN-only*.
  2. Гибридная (CNN + GNN + внимание) — *OCRModel*.
- Упрощённый визуальный энкодер (CNN) для ускорения.
- Обучение 90 эпох, батч 8, видеокарта P100.

Результаты

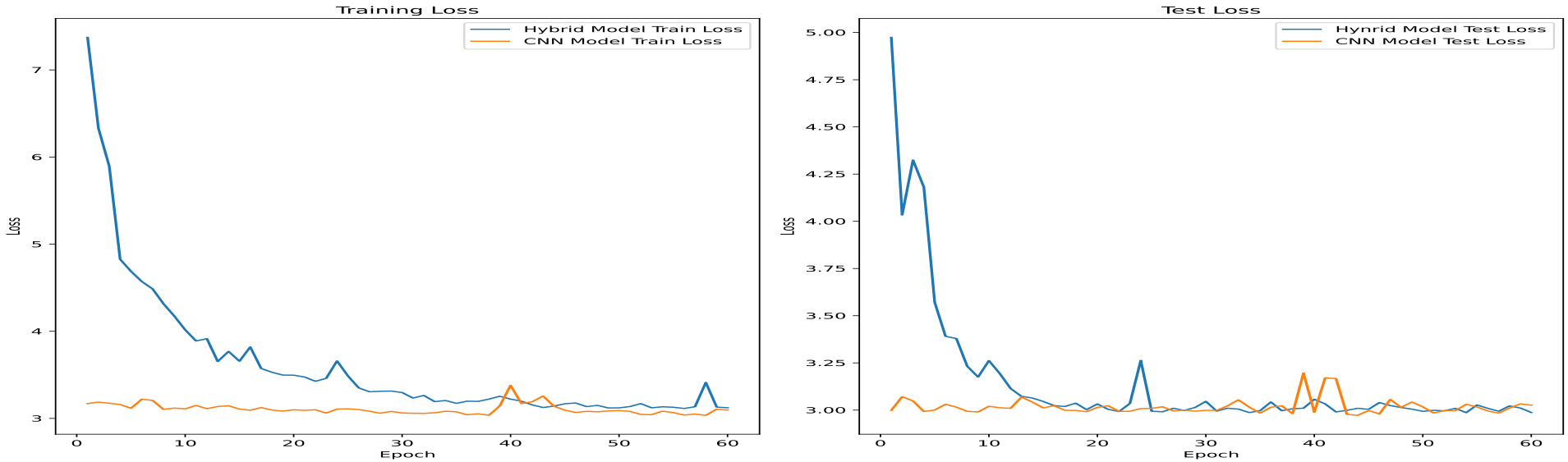


Рис. 4. Динамика функции потерь на обучающей и тестовой выборках (30–90 эпохи).

Модель	**Train Loss**	**Test Loss**
CNN-only	3.096	3.025
Гибридная	3.175	**3.009**

- Гибридная модель показывает **лучшее обобщение**.
- Графовый поток выступает в роли регуляризатора, снижая переобучение.

Выводы и перспективы

Основные результаты

- Предложена устойчивая гибридная архитектура с математически обоснованным построением графа штрихов.
- Экспериментально подтверждён регуляризующий эффект графового потока.
- Гибридная модель превосходит чисто визуальный аналог по обобщающей способности.

Направления дальнейших исследований

- Интеграция полноценного энкодера VAN.
- Расширение экспериментов на исторические коллекции с оценкой CER/WER.
- Исследование двунаправленного внимания.

Спасибо за внимание!

Готов ответить на ваши вопросы.

Литература

- Местецкий Л.М. Непрерывная морфология бинарных изображений. — М.: Физматлит // 2009.
- Mestetskiy L.M. Continuous Morphology of Binary Images: Figures, Skeletons, and Circular Arcs. — Springer // 2024.
- Зыков А.Г. и др. Vertical Attention Network для распознавания рукописного текста // 2024.
- Gan J. et al. GCM: Graph-Enhanced Cross-Modal Mutual Learning for Handwritten Text Recognition // 2025.
- Sharma D.K. et al. Graph Neural Networks for Handwriting Recognition // 2024.