Искусственный интеллект: эволюция идей от Фрэнсиса Бэкона до векторных трансформеров и ChatGPT

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН • профессор ВМК МГУ,
руководитель лаборатории машинного обучения
и семантического анализа Института ИИ МГУ,
г.н.с. ФИЦ ИУ РАН, профессор МФТИ

Институт искусственного интеллекта МГУ

Проблемы искусственного интеллекта — совместный научный семинар Российской ассоциации искусственного интеллекта и ФИЦ «Информатика и управление» РАН
• 19 апреля 2023 •

Содержание

Принцип эмпирической индукции

«Не следует полагаться на сформулированные аксиомы и формальные базовые понятия, какими бы привлекательными и справедливыми они не казались. Законы природы нужно «расшифровывать» из фактов опыта. Следует искать правильный метод анализа и обобщения опытных данных; здесь логика Аристотеля не подходит в силу её абстрактности, оторванности от реальных процессов и явлений.»

Фрэнсис Бэкон (1561–1626)

Таблицы открытия: множество случаев x, когда

- свойство y *присутствовало* $y(x) = 1$
- свойство y *отсутствовало* $y(x) = 0$
- наблюдалось изменение *степени* свойства $y(x)$

Фрэнсис Бэкон. Новый органон. 1620.

Метод наименьших квадратов (Гаусс, 1795)

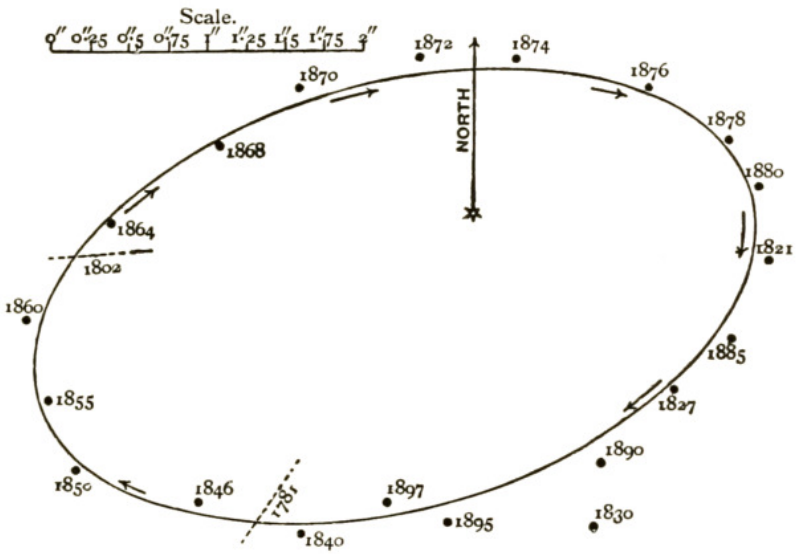
Линейная модель регрессии:

$$a(x, w) = \sum_{j=1}^{n} w_j f_j(x), \qquad w \in \mathbb{R}^n.$$

Метод наименьших квадратов:

$$Q(w) = \sum_{i=1}^{\ell} \big(a(x_i, w) - y_i\big)^2 \to \min_w .$$

Карл Фридрих Гаусс (1777–1855)



«Our principle, which we have made use of since 1795, has lately been published by Legendre...»

C.F.Gauss. Theory of the motion of the heavenly bodies moving about the Sun in conic sections. 1809.

Общая оптимизационная задача машинного обучения

Дано: обучающая выборка объектов $\{x_i\}_{i=1}^{\ell}$

Найти: вектор параметров w предсказательной модели $a(x, w)$

Критерий: минимум эмпирического риска

$$\sum_{i=1}^{\ell} L_i(w) \;\to\; \min_w$$

где $L_i(w)$ — функция потерь модели $a(x, w)$ на объекте x_i

Обобщение: минимум регуляризованного эмпирического риска

$$\sum_{i=1}^{\ell} L_i(w) + \sum_{j=1}^{r} \tau_j R_j(w) \;\to\; \min_w$$

где R_j — регуляризаторы, τ_j — коэффициенты регуляризации

Оптимизационная задача обучения модели регрессии

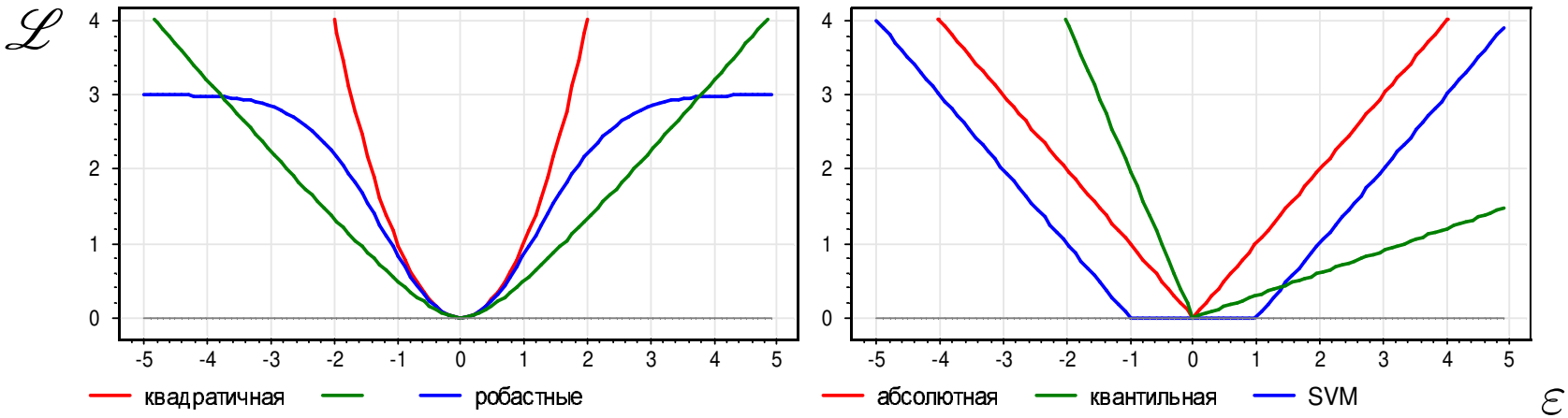
Дано: обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$ с ответами $y_i \in \mathbb{R}$

Найти: вектор параметров w модели регрессии $a(x, w)$

Критерий: минимум эмпирического риска

$$\sum_{i=1}^{\ell} \mathscr{L}\big(a(x_i, w) - y_i\big) \;\to\; \min_w$$

Унимодальные функции потерь $\mathscr{L}(\varepsilon)$ от невязки $\varepsilon = a(x, w) - y$:



Оптимизационная задача обучения модели классификации

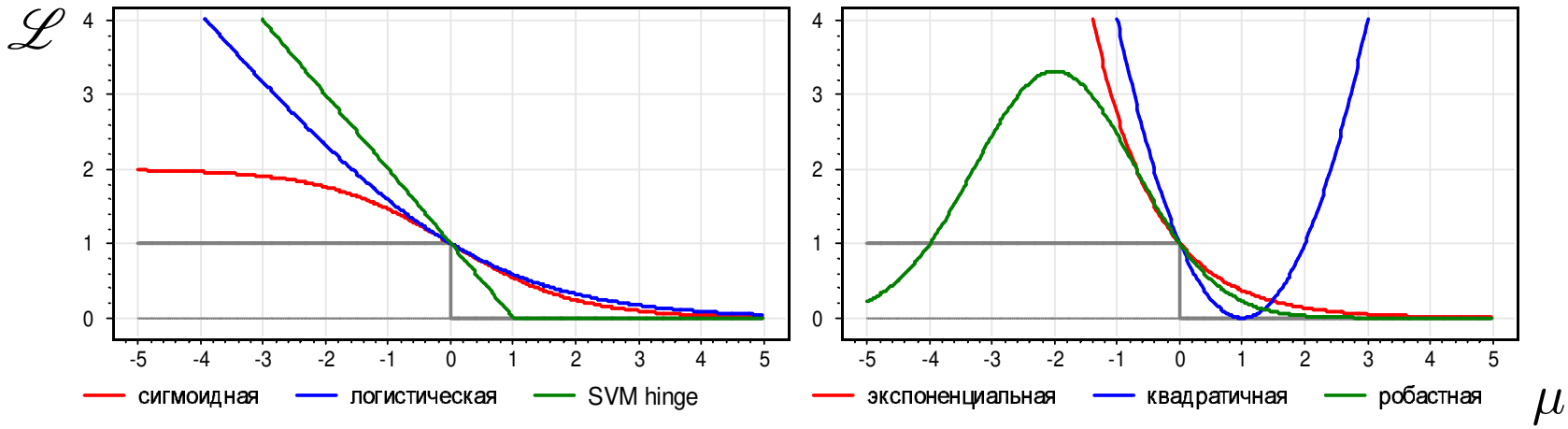
Дано: обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$, $y_i \in \{-1, +1\}$

Найти: вектор w модели классификации $a(x, w) = \operatorname{sign} g(x, w)$

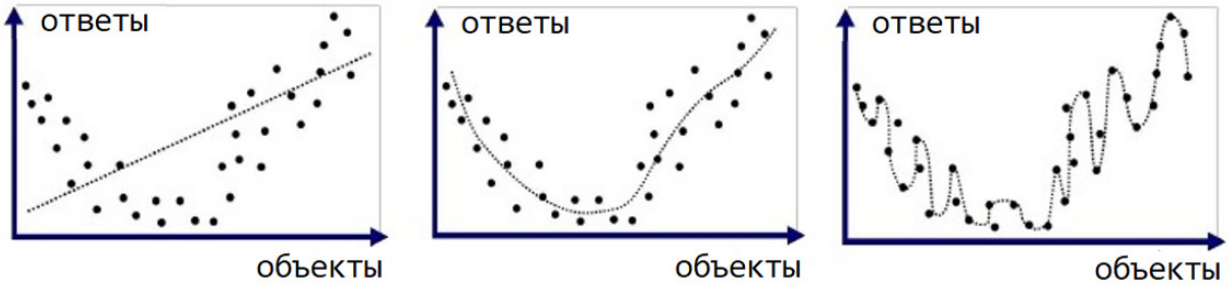
Критерий: *аппроксимация эмпирического риска*

$$\sum_{i=1}^{\ell} \big[g(x_i, w)y_i < 0\big] \leqslant \sum_{i=1}^{\ell} \mathscr{L}\big(g(x_i, w)y_i\big) \;\to\; \min_w$$

Убывающие функции потерь $\mathscr{L}(\mu)$ от отступа $\mu = g(x, w)y$:

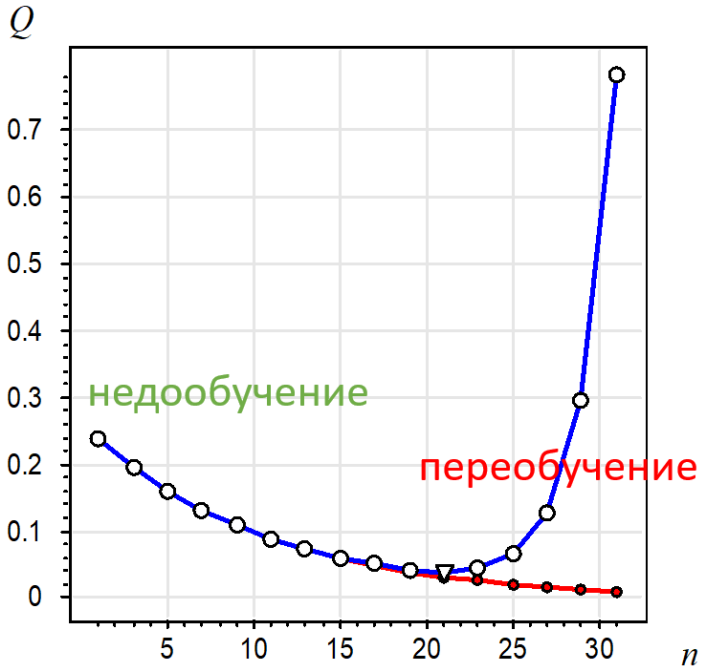


Проблемы недообучения и переобучения



недообучение переобучение

- Недообучение (underfitting): модель слишком проста, недостаточное число параметров n
- Переобучение (overfitting): модель слишком сложна, избыточное число параметров n



Понятие обучаемости в SLT, Statistical Learning Theory

Семейство классификаторов A *обучаемо*:

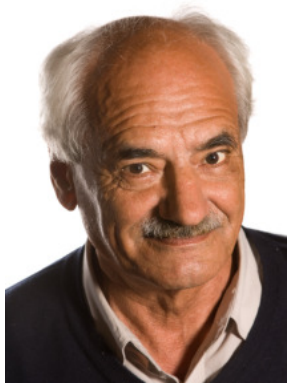
$$\mathsf{P}\left\{\sup_{a\in A}\left|P(a)-\nu(a,X^\ell)\right|>\varepsilon\right\}\leqslant\eta,$$

$P(a)$ — вероятность ошибки классификатора,
$\nu(a,X^\ell)$ — эмпирический риск (частота ошибок классификатора a на выборке).

Основные результаты VC-теории:

- Обосновано ограничение сложности A
- Понятие *ёмкости* семейства, VCdim
- Метод структурной минимизации риска

Владимир Наумович Вапник



Алексей Яковлевич Червоненкис (1938–2014)

Вапник В. Н., Червоненкис А. Я.
Теория распознавания образов. М.: Наука, 1974.

От простых регуляризаций к обучению сложных моделей

- LASSO (least absolute shrinkage and selection operator)

  Tibshirani R. Regression shrinkage and selection via the LASSO. 1996

- ElasticNet (сумма L_0 и L_1 регуляризаторов)

  Hui Zou, Hastie T. Regularization and variable selection via the Elastic Net. 2005

- Негладкие регуряризаторы для отбора признаков

  Tatarchuk A., Urlov E., Mottl V., Windridge D. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. 2010.

- Ансамблирование, голосование, комитеты, бустинг

  Мазуров В. Д. Комитеты системы неравенств и задача распознавания. 1971.
  Журавлёв Ю. И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. 1977.
  Freund Y., Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting. 1995.

- Dropout для глубоких нейронных сетей

  N.Srivastava, G.Hinton, A.Krizhevsky, I.Sutskever, R.Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. 2014

- Остаточные нейронные сети (Residual NN)

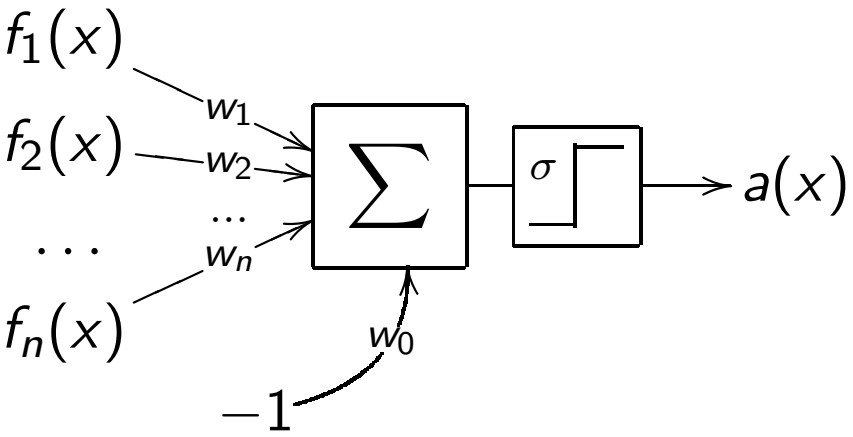
  Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. 2015.

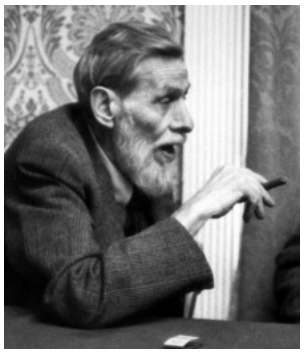
Искусственный нейрон — линейная модель классификации

Линейная модель нейрона (1943):

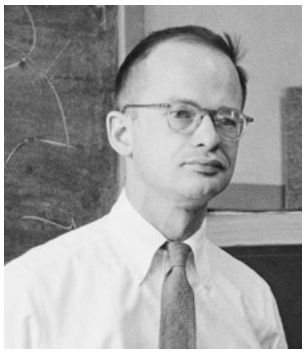
$$a(x,w) = \sigma\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right)$$

$f_j(x)$ — признаки объекта x
w_j — веса признаков
w_0 — порог активации
$\sigma(z)$ — функция активации

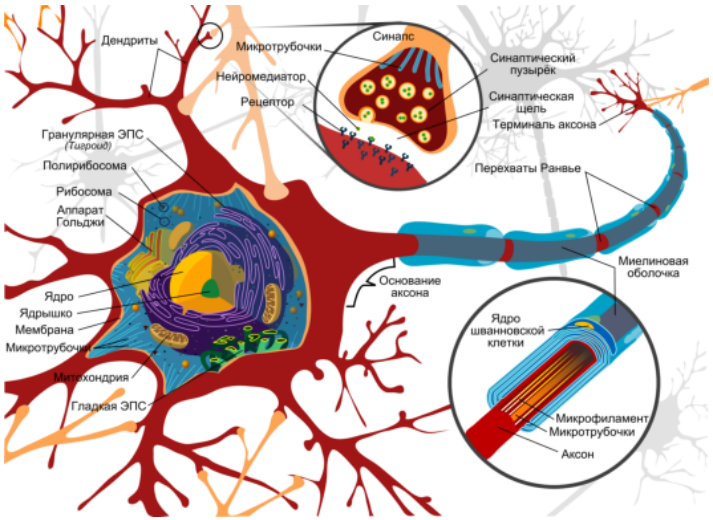




Уоррен МакКаллок (1898–1969)

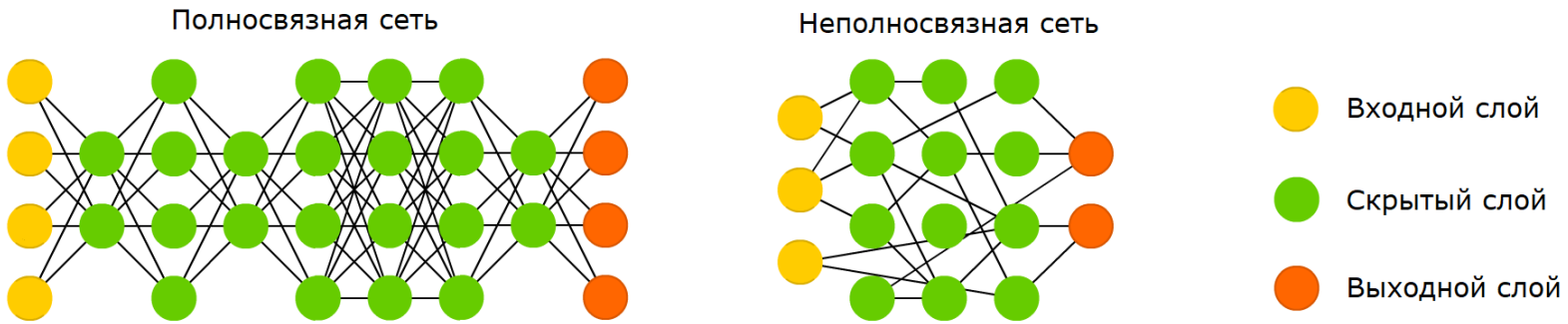


Вальтер Питтс (1923–1969)



Глубокие нейронные сети (Deep Neural Network, DNN)

1965: первые глубокие нейронные сети
1997: рекуррентная сеть LSTM для анализа последовательностей
2012: свёрточная сеть для классификации изображений AlexNet



- *Архитектура сети* — структура слоёв и связей между ними, позволяющая наделять DNN нужными свойствами
- DNN позволяют принимать на входе и генерировать на выходе *сложно структурированные данные*

Ива́хненко А. Г., Лапа В. Г. Кибернетические предсказывающие устройства. 1965
Hochreiter S., Schmidhuber J. Neural Computation, 9(8), 1997
Krizhevsky A. et al. ImageNet classification with deep convolutional neural networks. 2012

Генерация признаков для распознавания изображений

Классический подход к распознаванию изображений:

Конструируемые признаки изображения → **Любой обучаемый классификатор**

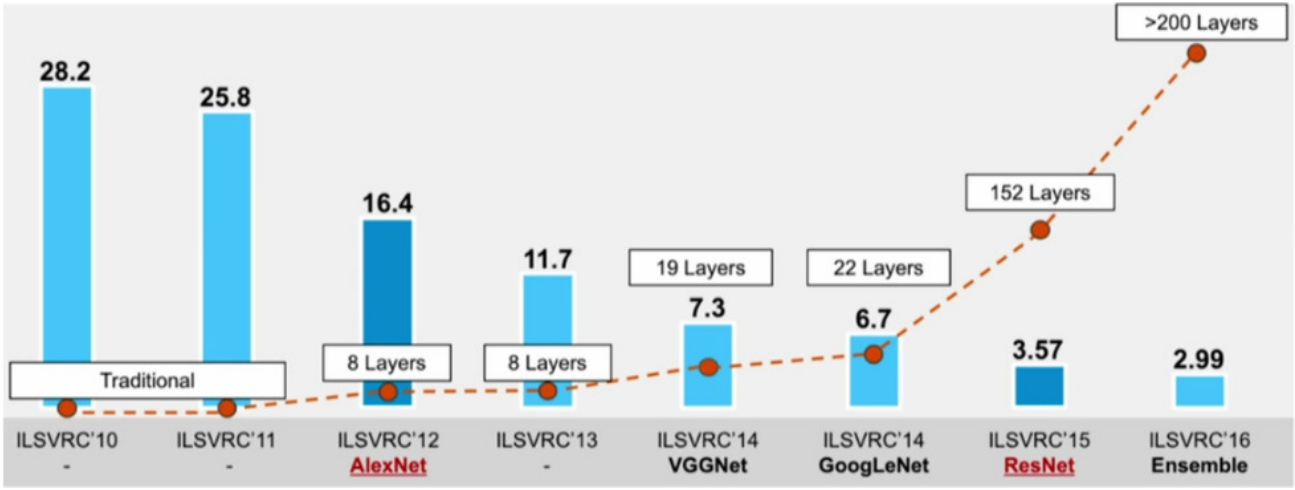
Современный подход — end-to-end Deep Learning:

Обучаемые ANN признаки изображения → **Обучаемый ANN классификатор**

Совместное обучение

Sanjeev Arora. Toward theoretical understanding of deep learning. ICML-2018 Tutorial
https://unsupervised.cs.princeton.edu/deeplearningtutorial.html

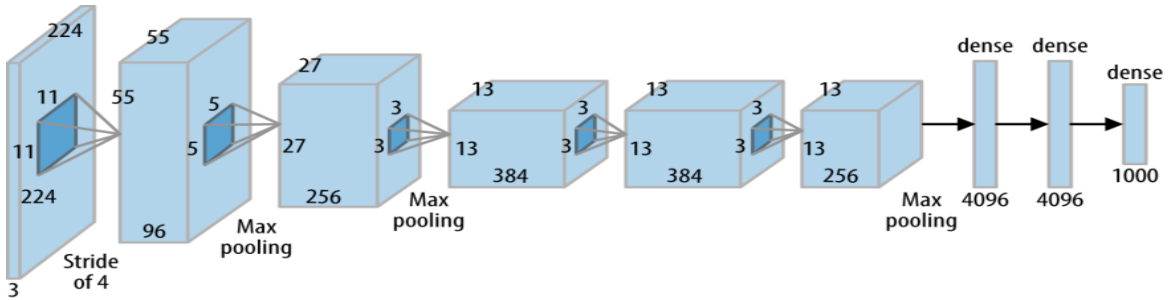
Глубокие свёрточные сети для классификации изображений



Старт в 2009

Человеческий уровень ошибок 5% пройден в 2015

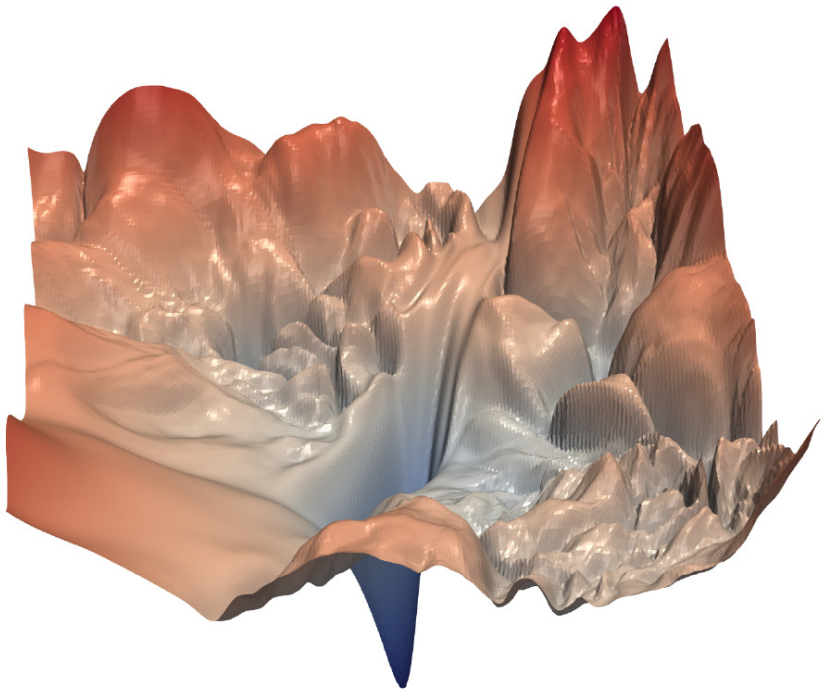
Свёрточные нейронные сети
AlexNet (2012)
ResNet (2015)



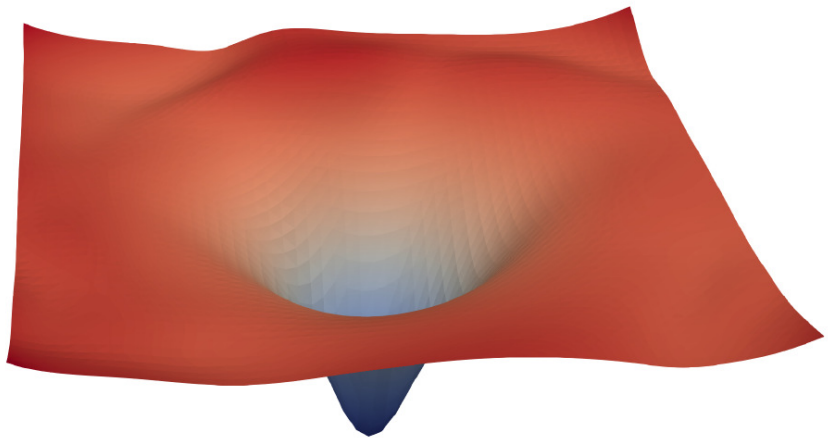
Li Fei-Fei et al. ImageNet: A large-scale hierarchical image database. 2009
Krizhevsky A. et al. ImageNet classification with deep convolutional neural networks. 2012
Kaiming He et al. Deep residual learning for image recognition. 2015

ResNet: визуализация оптимизационного критерия

Сквозные связи (skip connection) упрощают оптимизируемый критерий, устраняя локальные экстремумы и седловые точки:



without skip connections



with skip connections

Hao Li et al. Visualizing the Loss Landscape of Neural Nets. 2018

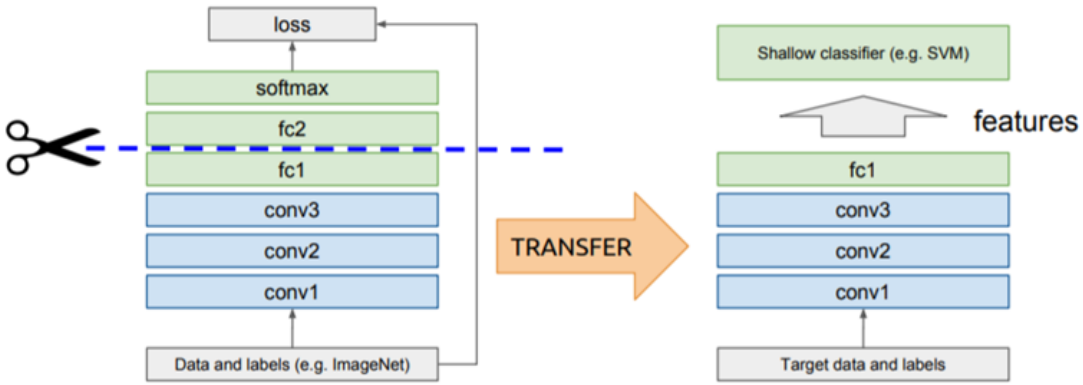
Предобучение (pre-training), перенос обучения (transfer learning)

Обучение модели векторизации $z = f(x, \alpha)$ на выборке $\{x_i\}_{i=1}^{\ell}$:

$$\sum_{i=1}^{\ell} \mathscr{L}_i\big(g(f(x_i, \alpha), \beta)\big) \;\to\; \min_{\alpha,\beta}$$

Обучение целевой модели $y = g(z, \beta)$ на малых данных:

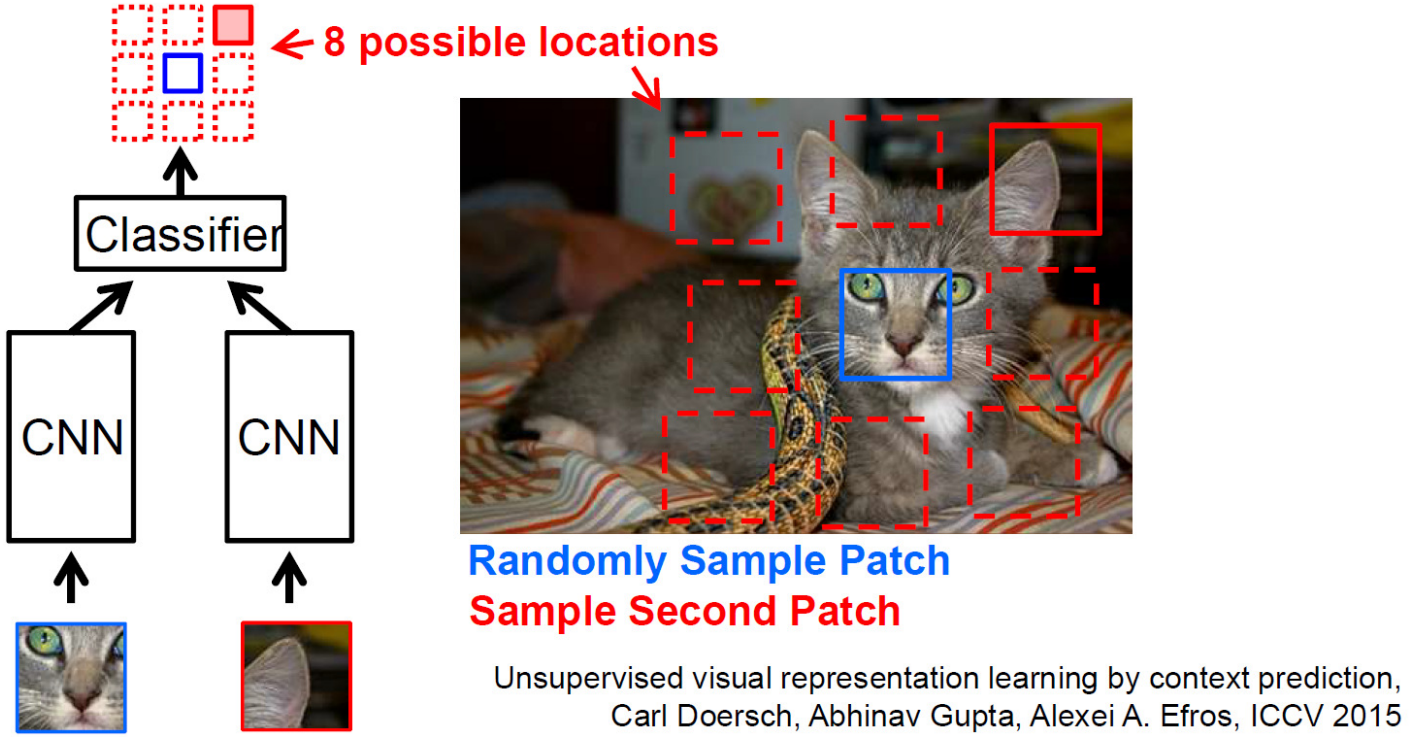
$$\sum_{i=1}^{m} \mathscr{L}'_i\big(g'(f(x'_i, \alpha), \beta')\big) \;\to\; \min_{\beta'}$$



Sinno Jialin Pan, Qiang Yang. A Survey on Transfer Learning. 2009
J.Yosinski et al. How transferable are features in deep neural networks? 2014.

Самостоятельное обучение (self-supervised learning)

Модель векторизации $z = f(x, \alpha)$ обучается предсказывать взаимное расположение пар фрагментов одного изображения



Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Преимущество: сеть выучивает векторные представления объектов без размеченной обучающей выборки (без ImageNet).

Многозадачное обучение (multi-task learning)

$z = f(x, \alpha)$ — векторизация, универсальная для всех моделей
$g_t(z, \beta)$ — специфичная часть модели для задачи $t \in T$

Одновременное обучение модели f по задачам X_t, $t \in T$:

$$\sum_{t \in T} \sum_{i \in X_t} \mathscr{L}_{ti}\big(g_t(f(x_{ti}, \alpha), \beta_t)\big) \;\to\; \min_{\alpha, \{\beta_t\}}$$

Обучаемость (learnability): качество решения отдельной задачи $\langle X_t, \mathscr{L}_t, g_t \rangle$ улучшается с ростом объёма выборки $\ell_t = |X_t|$.

Learning to learn: качество решения каждой из задач $t \in T$ улучшается с ростом как ℓ_t, так и общего числа задач $|T|$.

Few-shot learning: для решения новой задачи t достаточно небольшого числа примеров, иногда даже одного.

M.Crawshaw. Multi-task learning with deep neural networks: a survey. 2020
Y.Wang et al. Generalizing from a few examples: a survey on few-shot learning. 2020

Обучаемая векторизация данных: задача автокодировщика

Дано: обучающая выборка объектов $\{x_i\}_{i=1}^{\ell}$

Найти: $z = f(x, \alpha)$ — модель кодировщика (encoder)
$\hat{x} = g(z, \beta)$ — модель декодировщика (decoder)

Критерий: качество реконструкции исходных объектов

$$\sum_{i=1}^{\ell} \mathscr{L}\big(g(f(x_i, \alpha), \beta), x_i\big) \;\to\; \min_{\alpha,\beta}$$

Квадратичная функция потерь: $\mathscr{L}(\hat{x}, x) = \|\hat{x} - x\|^2$

Для *линейного автокодировщика* $f(x, A) = Ax$, $g(z, B) = Bz$,
задача сводится к (низкоранговому) матричному разложению:

$$\sum_{i=1}^{\ell} \|BAx_i - x_i\|^2 \to \min_{A,B}$$

Векторизация совместно с предсказательным моделированием

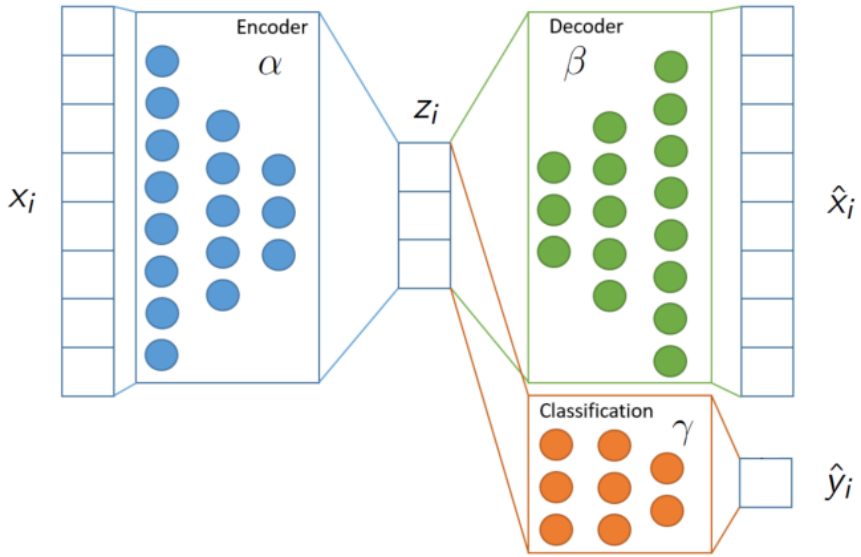
Данные: размеченные $(x_i, y_i)_{i=1}^{k}$, неразмеченные $(x_i)_{i=k+1}^{\ell}$

Найти:

$z_i = f(x_i, \alpha)$ — кодировщик

$\hat{x}_i = g(z_i, \beta)$ — декодировщик

$\hat{y}_i = \hat{y}(z_i, \gamma)$ — предиктор



Функции потерь:

$\mathscr{L}(\hat{x}_i, x_i)$ — реконструкция

$\tilde{\mathscr{L}}(\hat{y}_i, y_i)$ — предсказание

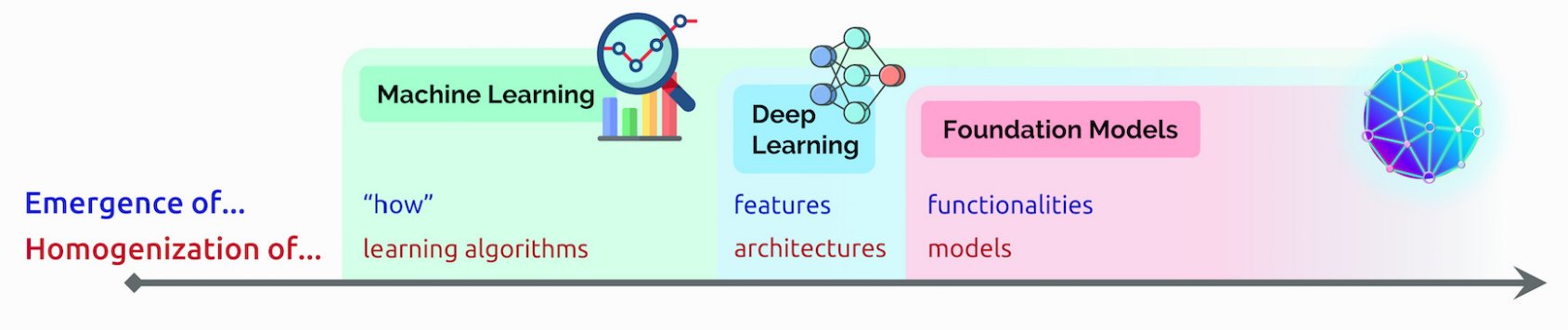
Критерий: совместное обучение автокодировщика и предсказательной модели (классификации, регрессии или др.):

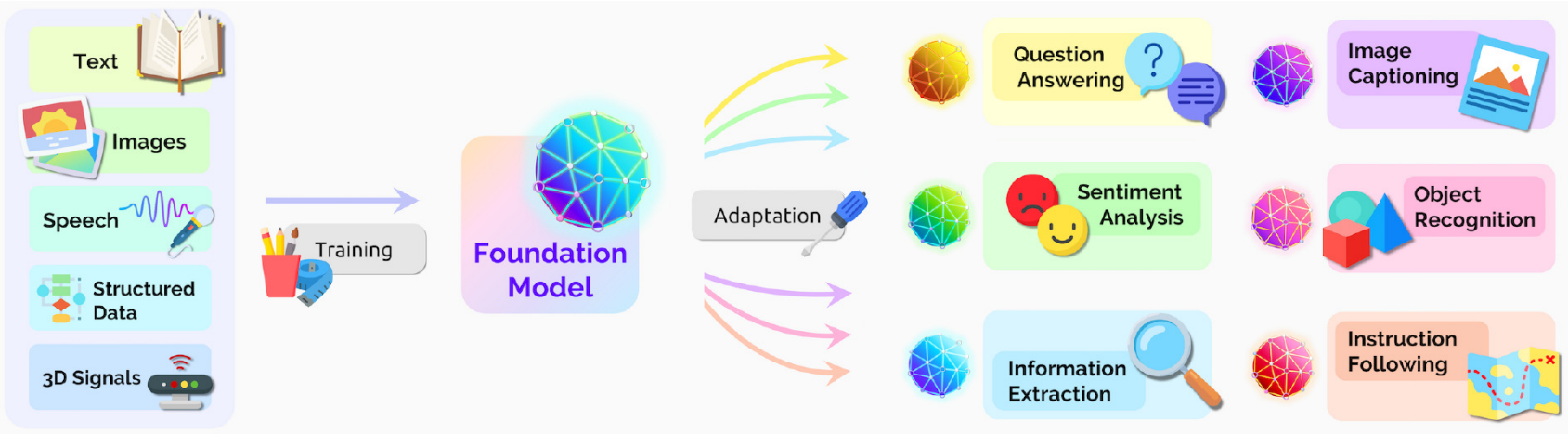
$$\sum_{i=1}^{\ell} \mathscr{L}\big(g(f(x_i, \alpha), \beta), x_i\big) + \lambda \sum_{i=1}^{k} \tilde{\mathscr{L}}(\hat{y}(f(x_i, \alpha), \gamma), y_i) \;\to\; \min_{\alpha,\beta,\gamma}$$

Dor Bank, Noam Koenigstein, Raja Giryes. Autoencoders. 2020

Обучаемая векторизация данных — глобальный тренд AI/ML

Foundation Models — гомогенизация векторных моделей





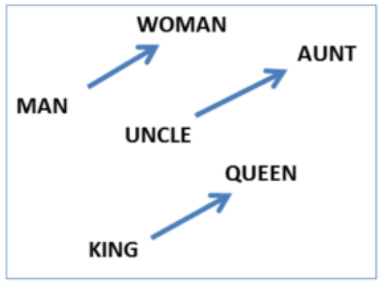
R.Bommasani et al. (Center for Research on Foundation Models, Stanford University) On the opportunities and risks of foundation models // CoRR, 20 August 2021.

Эволюция подходов машинного обучения в анализе текстов

Декомпозиция задач по уровням пирамиды NLP

- морфологический анализ, лемматизация, опечатки
- синтаксический анализ, выделение терминов, NER
- семантический анализ, выделение фактов, тем

Модели векторных представлений (эмбедингов) слов на основе матричных разложений



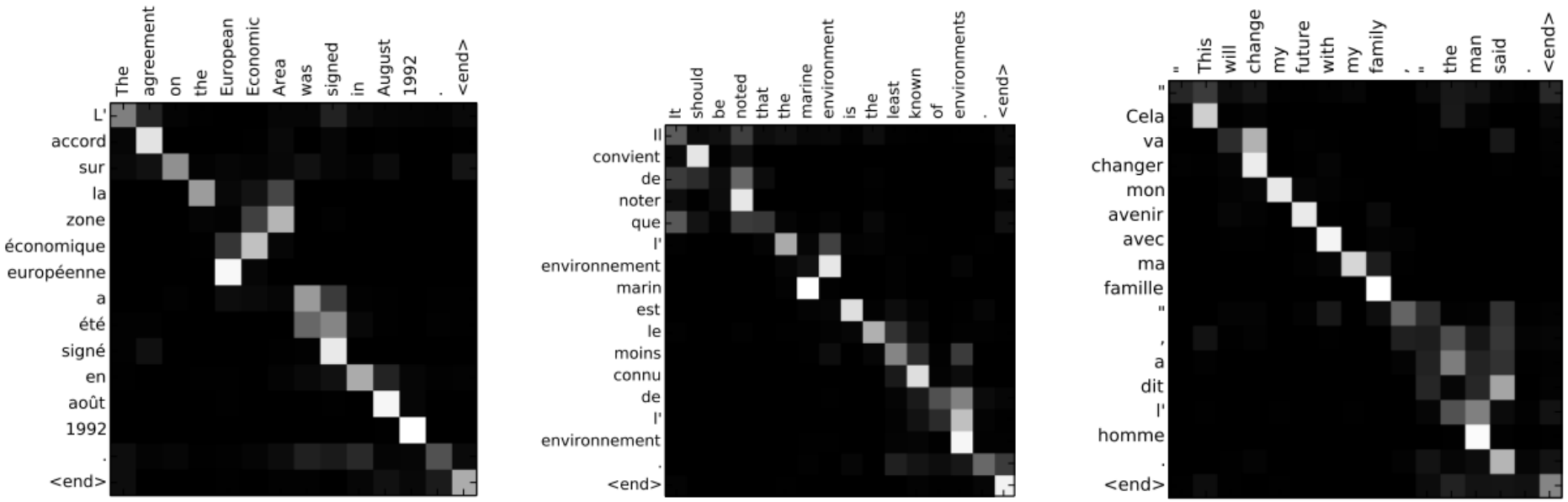
- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016]
- тематические модели LDA [Blei, 2003], ARTM [2014]

Нейросетевые модели локальных контекстов

- рекуррентные нейронные сети
- модели внимания и трансформеры:
  BERT [2018], GPT-3 [2020], GPT-4 [2023]

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d}}\right) V$$

Модели внимания для машинного перевода



Вход: $\{x_i\}$ — последовательность слов входного языка
Выход: $\{y_t\}$ — последовательность слов выходного языка

Интерпретация: матрица a_{it} показывает, на какие слова x_i модель обращает внимание, генерируя слово перевода y_t

---

Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015.

Модели внимания для аннотирования изображений

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Подсвечены области, на которые модель обращает внимание, когда генерирует подчёркнутое слово в аннотации изображения

Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

Трасформер для машинного перевода

Трасформер (transformer) — это нейросетевая архитектура на основе моделей внимания и полносвязных слоёв

Схема преобразований данных в машинном переводе:

- $S = (w_1, \ldots, w_n)$ — слова предложения на входном языке
  $\downarrow$ обучаемая или пред-обученная векторизация слов
- $X = (x_1, \ldots, x_n)$ — векторы слов входного предложения
  $\downarrow$ трансформер-кодировщик
- $Z = (z_1, \ldots, z_n)$ — контекстные векторы слов
  $\downarrow$ трансформер-декодировщик, похож на кодировщика
- $Y = (y_1, \ldots, y_m)$ — векторы слов выходного предложения
  $\downarrow$ генерация слов из построенной языковой модели
- $\tilde{S} = (\tilde{w}_1, \ldots, \tilde{w}_m)$ — слова предложения на выходном языке

Vaswani et al. (Google) Attention is all you need. 2017.

Модель внимания Query–Key–Value

q — *вектор-запрос* для трансформации в вектор-контекст z
$K = (k_1, \ldots, k_n)$ — *векторы-ключи*, сравниваемые с запросом
$V = (v_1, \ldots, v_n)$ — *векторы-значения*, образующие контекст

Модель внимания — трёхслойная сеть, вычисляющая z как выпуклую комбинацию векторов v_i, релевантных запросу q:

$$z = \text{Attn}(q, K, V) = \sum_i v_i \, \text{SoftMax}_i \langle k_i, q \rangle,$$

где $\langle k_i, q \rangle$ — *оценка релевантности* ключа k запросу q

Модель внутреннего внимания (самовнимания, self-attention):

$$z_i = \text{Attn}(W_q x_i, W_k X, W_v X)$$

трансформирует входную последовательность $X = (x_1, \ldots, x_n)$
в выходную последовательность векторов контекста $(z_1, \ldots, z_n)$

Архитектура трансформера-кодировщика

1. Добавляются позиционные векторы p_i:
$h_i = x_i + p_i, \quad H = (h_1, \ldots, h_n)$
$d = \dim x_i, p_i, h_i = 512$
$\dim H = 512 \times n$

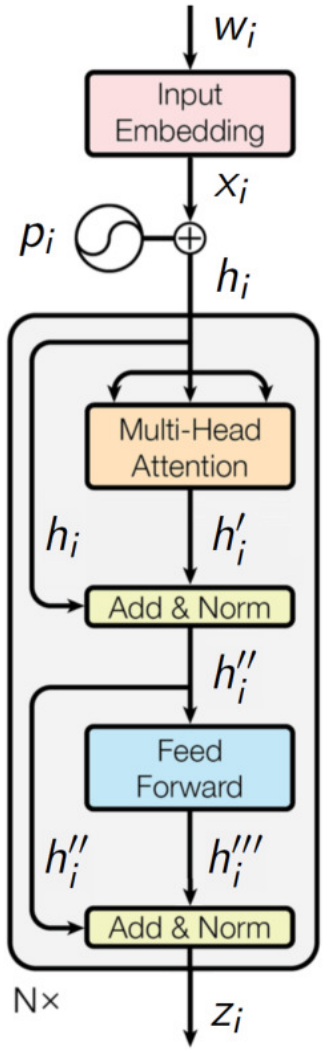
2. Многомерное самовнимание:
$h_i^j = \text{Attn}(W_q^j h_i, W_k^j H, W_v^j H)$
$j = 1, \ldots, J = 8$
$\dim h_i^j = 64$
$\dim W_q^j, W_k^j, W_v^j = 64 \times 512$

3. Конкатенация:
$h'_i = \text{MH}_j(h_i^j) \equiv \left[h_i^1 \cdots h_i^J\right]$
$\dim h'_i = 512$

4. Сквозная связь + нормировка уровня:
$h''_i = \text{LN}(h'_i + h_i; \mu_1, \sigma_1)$
$\dim h''_i, \mu_1, \sigma_1 = 512$

5. Полносвязная 2х-слойная сеть FFN:
$h'''_i = W_2\,\text{ReLU}(W_1 h''_i + b_1) + b_2$
$\dim W_1 = 2048 \times 512$
$\dim W_2 = 512 \times 2048$

6. Сквозная связь + нормировка уровня:
$z_i = \text{LN}(h'''_i + h''_i; \mu_2, \sigma_2)$
$\dim z_i, \mu_2, \sigma_2 = 512$

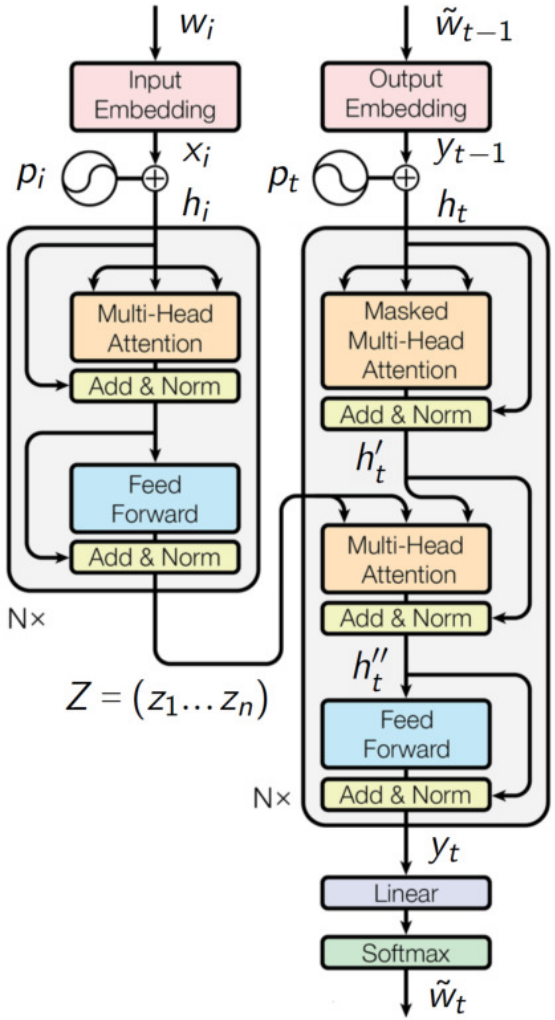


Архитектура трансформера декодировщика

Авторегрессионный синтез последовательности:
$y_0 = \langle \text{BOS} \rangle$ — эмбеддинг символа начала;
для всех $t = 1, 2, \ldots$:

1. Маскирование «данных из будущего»:
   $h_t = y_{t-1} + p_t; \quad H_t = (h_1, \ldots, h_t)$
2. Многомерное самовнимание:
   $h'_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(W^j_q h_t, W^j_k H_t, W^j_v H_t)$
3. Многомерное внимание на кодировку Z:
   $h''_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(\tilde{W}^j_q h'_t, \tilde{W}^j_k Z, \tilde{W}^j_v Z)$
4. Двухслойная полносвязная сеть:
   $y_t = \text{LN} \circ \text{FFN}(h''_t)$
5. Линейный предсказывающий слой:
   $p(\tilde{w}|t) = \text{SoftMax}_{\tilde{w}}(W_y y_t + b_y)$

генерация $\tilde{w}_t = \arg\max_{\tilde{w}} p(\tilde{w}|t)$ **пока** $\tilde{w}_t \neq \langle \text{EOS} \rangle$



Vaswani et al. (Google) Attention is all you need. 2017.

BERT (Bidirectional Encoder Representations from Transformers)

Трансформер BERT — это кодировщик без декодировщика, предобучаемый на большой текстовой коллекции для решения широкого класса задач NLP

Схема преобразования данных в задачах NLP:

- $S = (w_1, \ldots, w_n)$ — токены предложения входного текста
  $\downarrow$ обучение эмбедингов вместе с трансформером
- $X = (x_1, \ldots, x_n)$ — эмбединги токенов входного предложения
  $\downarrow$ трансформер кодировщика
- $Z = (z_1, \ldots, z_n)$ — трансформированные эмбединги
  $\downarrow$ дообучение на конкретную задачу
- Y — выходной текст / разметка / классификация и т.п.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language) BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Критерии обучения трансформеров

- **Машинный перевод:** максимизация правдоподобия слов перевода $\tilde{w}_t$ по выборке пар предложений «S, перевод $\tilde{S}$»:

$$\sum_{(S,\tilde{S})} \sum_{\tilde{w}_t \in \tilde{S}} \ln p(\tilde{w}_t | t, S, W) \to \max_W$$

- **BERT MLM (masked language modeling):** предсказание пропущенных слов по локальному контексту
- **BERT NSP (next sentence prediction):** предсказание, следуют ли два предложения друг за другом
- **Fine-tuning:** дообучение трансформера $Z(S, W)$ на задаче с моделью $f(Z(S, W), W_f)$, выборкой $\{S\}$ и $\mathscr{L}(S, f) \to \max$
- **Multi-task learning:** дообучение на наборе задач $\{t\}$ с моделями $f_t(Z(S, W), W_t)$, выборками $\{S\}_t$, по сумме критериев $\sum_t \lambda_t \sum_S \mathscr{L}_t(S, f_t) \to \max$

ChatGPT и GPT-4: проблески общего искусственного интеллекта

Sparks of Artificial General Intelligence: Early experiments with GPT-4

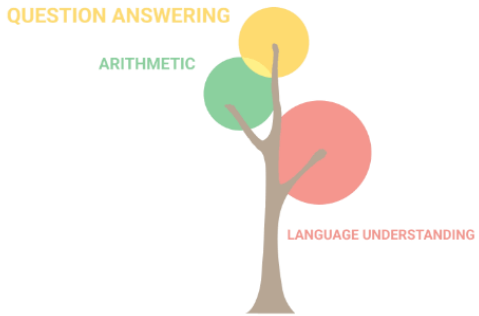
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research *(27 March 2023)*

Новые способности модели, не закладывавшиеся при обучении:

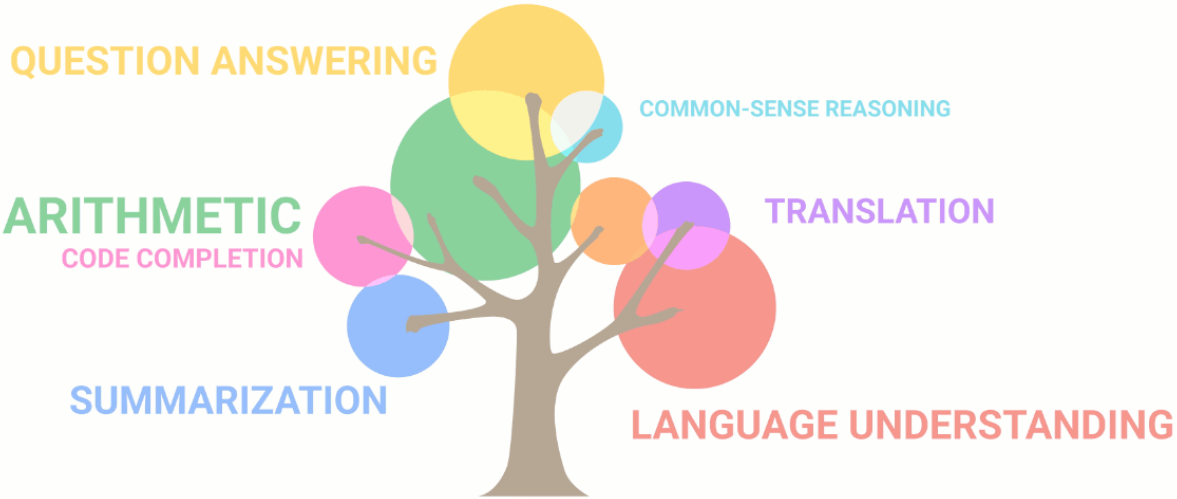
- объяснять свои ответы, перефразировать
- реферировать, генерировать планы, сценарии, шаблоны
- переводить на другие языки, строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать и исправлять собственные ошибки по подсказке

Появление у модели качественно новых способностей



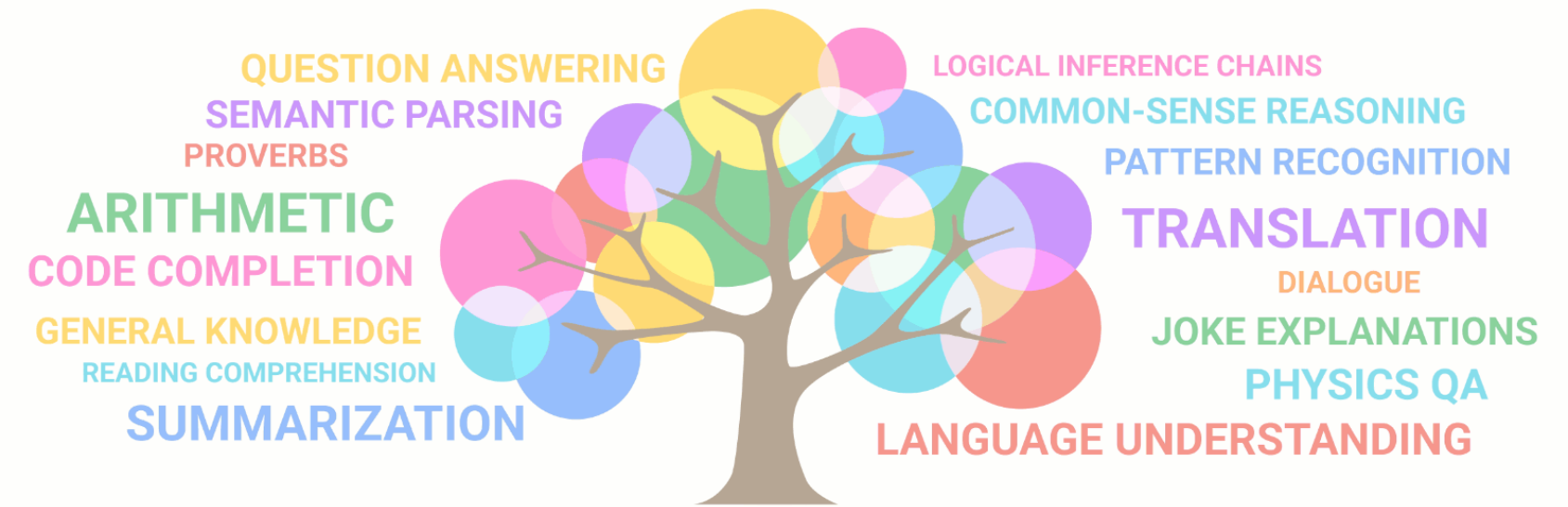
- GPT-2: 14/Feb/2019, контекст 768 слов (1,5 страницы)
- 1,5 млрд. параметров, корпус 10 млрд. токенов (40Gb)
- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

Появление у модели качественно новых способностей



- GPT-3: 11/Jun/2020, контекст 1536 слов (3 страницы)
- 175 млрд. параметров, корпус 500 млрд. токенов
- способность делать перевод на другие языки,
- решать логические и математические задачи,
- генерировать программный код по описанию

Появление у модели качественно новых способностей



- GPT-4: 14/Mar/2023, контекст 24 000 слов (48 страниц)
- >1 трл. параметров, корпус >1Tb
- способность описывать и анализировать изображения,
- реагировать на подсказки вроде «Let's think step by step»,
- решать качественные физические задачи по картинке

Выводы по части 1 (из 3)

Основные принципы и составные части успеха ИИ-технологий:

- принцип эмпирической индукции Фрэнсиса Бэкона
- минимизация (и аппроксимация) эмпирического риска
- регуляризация некорректно поставленных задач
- коннекционизм и глубокие нейросетевые архитектуры
- градиентная оптимизация сверхвысокой размерности
- векторизация сложно структурированных данных
- модели внимания для учёта данных контекста
- самостоятельное обучение вместо обучения по разметке
- увеличение объёмов (не)размеченных данных
- увеличение размера моделей при контроле переобучения
- увеличение скорости и параллелизма вычислителей

Далеко не полный перечень возможностей

Персональный интеллектуальный помощник — «бот»
уже способен помогать с рутинно-творческой работой:

- делать обзоры, рефераты, сводки на разных языках
- искать и структурировать профессиональную информацию
- сообщать новости, поддерживать разговор по теме
- генерировать документы или сайты по описанию
- в том числе юридические документы по шаблонам
- генерировать программный код по описанию
- уточнять и дополнять контент по просьбе, в диалоге
- разговаривать с детьми с учётом возрастных особенностей
- выполнять функции воспитателя, учителя, наставника
- оказывать психологическую помощь . . .

Далеко не полный перечень угроз

Персональный интеллектуальный помощник — «бот»
способен представлять угрозу, даже не обладая автономностью:

- «галлюцинируя», сообщать неверную информацию
- давать неверные сведения, касающиеся здоровья человека,
- других людей, событий, технологий, норм, правил, законов
- вызывать необоснованное доверие и манипулировать
- побуждать человека к действиям, не выгодным ему
- побуждать изменить точку зрения, умалчивая информацию
- поддерживать предрассудки и лженаучные представления
- поддерживать пропагандистские медиа-кампании
- влиять на формирование мировоззрения детей и подростков
- оказывать депрессивное воздействие на психику ...

Далеко не полный перечень задач и мер безопасности

Персональный интеллектуальный помощник — «бот» должен обладать способностями, явно закладываемыми в модель:

- *быть более экстрактивным*: подкреплять сведение или точку зрения дословной цитатой со ссылкой на источник
- *не занимать одну из сторон в человеческих конфликтах:* детектировать поляризацию, излагать позиции всех сторон
- *не имитировать* взятие ответственности на себя, наличие собственного мнения, личности, какой-либо субъектности
- *не пытаться заменять решение задачи рассуждением*, уметь взаимодействовать с внешними решателями
- решать *трудные задачи понимания естественного языка*: уметь объяснять контекст, подтекст, затекст, интертекст с явной опорой на обширные гуманитарные знания . . .

Пример 1. Конкурс ПРО//ЧТЕНИЕ

Задача: поиск смысловых ошибок в сочинениях ЕГЭ по русскому, литературе, истории, обществознанию, английскому

Период: декабрь 2019 — декабрь 2022

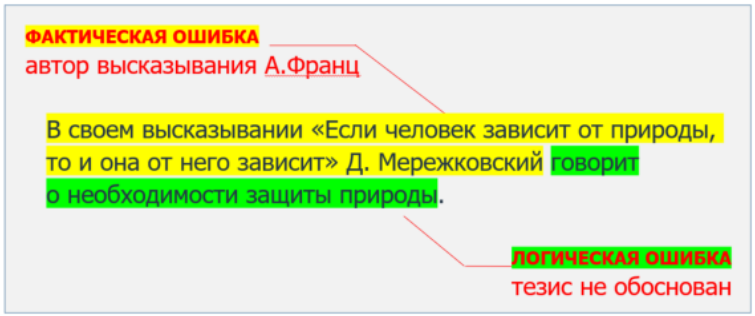
Призовой фонд:
— 100M руб. русский язык
— 100M руб. английский язык

Up Great technology contest

READ // ABLE
comprehension technology

Типов ошибок: 152
(р:70 л:16 о:23 и:20 а:23)
Подтипов ошибок: 236
(р:112 л:19 о:29 и:26 а:50)

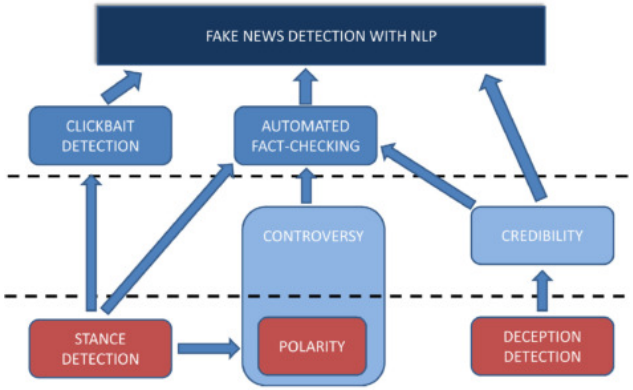
Алгоритм должен выделять ошибок и давать их объяснения.



Технический регламент конкурса ПРО//ЧТЕНИЕ: http://ai.upgreat.one. 2019–2022

Пример 2. Область исследований «Fake News Detection»

1. **Deception Detection**
   выявление обмана в тексте
2. **Automated Fact-Checking**
   автоматическая проверка фактов
3. **Stance Detection**
   выявление позиции за или против
4. **Controversy Detection**
   выявление и кластеризация разногласий
5. **Polarization Detection**
   выявление полярных позиций
6. **Clickbait Detection**
   противоречия заголовка и текста
7. **Credibility Scores**
   оценка достоверности источников



E.Saquete et al.
Fighting post-truth using natural language processing: a review and open challenges // Expert Systems With Applications, Elsevier, 2020.

Типология угроз и задачи их автоматической детекции

воздействия → фейки → пропаганда → инф.война

1. ■□■■ детекция приёмов манипулирования
2. ■□■■ детекция замалчивания
3. □■■■ детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)
4. □■□□ детекция кликбэйта (clickbait detection)
5. □■□□ автоматическая проверка фактов (auto fact-checking)
6. □■■□ детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)
7. □□■■ выявление конструктов картины мира: идеологем, мифологем
8. ■□■■ оценивание возможных психо-эмоциональных реакций
9. ■□■■ выявление целевых аудиторий воздействия
10. □■■■ оценивание и предсказание скорости распространения (virality prediction)
11. □■■□ оценивание достоверности источников (credibility scores)
12. □□□■ детекция прямой агрессии (угрозы, призывы, провокации, вербовка, экстремизм)

E.Saquete, D.Tomas, P.Moreda, P.Martinez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges // Expert Systems With Applications, Elsevier, 2020.

Типы задач ML/NLU для мониторинга медиа-пространства

1. **Классификация текста (сообщения/предложения) целиком**
   - deception detection, fact-checking, text credibility
2. **Классификация пары текстов**
   - stance, controversy, polarization, clickbait detection
   - выявление противоречий, разногласий, замалчивания
3. **Разметка текста (выделение и классификация фрагментов)**
   - поиск лингвистических маркеров (linguistic-based cues) в тексте
   - детекция приёмов манипулирования
   - выявление конструктов картины мира: мифологем, идеологем
   - выявление психо-эмоциональных реакций и целевых аудиторий
4. **Кластеризация или тематическое моделирование**
   - кластеризация мнений по заданной теме (controversy detection)
   - выявление поляризованных мнений (polarization detection)
   - выявление мнений как сочетаний слов, семантических ролей и тональностей
   - выявление «картин мира» – устойчивых сочетаний суждений и идеологем

Задачи Propaganda/Manipulation/Persuasion Detection

Базовая разметка: «фрагмент, метка класса»

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language
Attack on Reputation: Smears
Manipulative Wording: Exaggeration
Justification: Appeal to Values

Commissio
PopulusQue
Europaea

Упрощённая разметка: «предложение, метка класса»

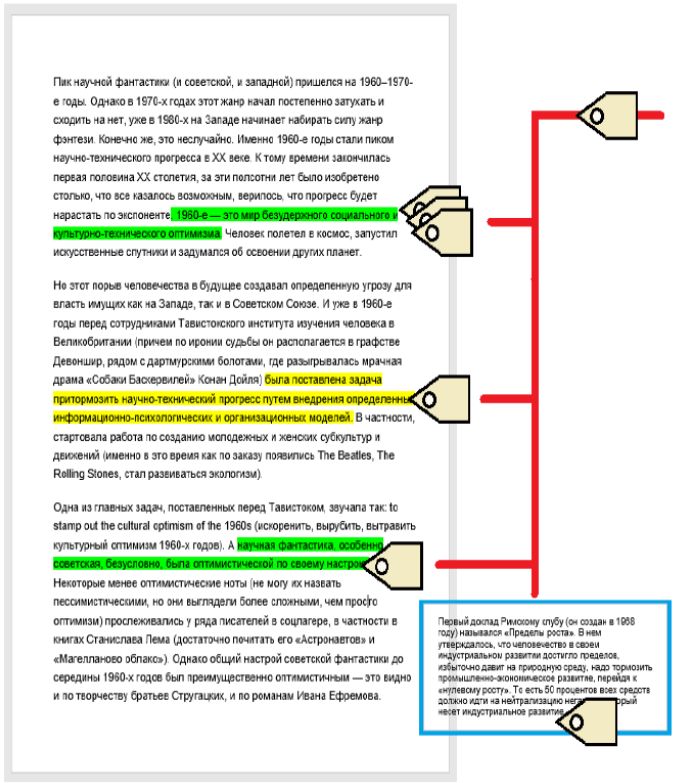
Продвинутая разметка: «фрагмент, мишень, метка класса»

SemEval-2023 task 3. Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup.
https://propaganda.math.unipd.it/semeval2023task3

G.Martino, P.Nakov et al. A survey on computational propaganda detection. 2020.

Унификация профессиональной разметки

Обобщение классических задач компьютерной лингвистики (NER, SemAn, SemRL, SyntPars), задач выявления манипуляций, поляризации, смысловых ошибок в академических эссе и др.



Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, затекстов и тегов

Теги (классы) выбираются из словаря тегов

Фрагмент задаётся началом и концом, может иметь один или несколько тегов:

BEGIN была поставлена задача притормозить научно-технический прогресс путем внедрения определен информационно-психологических и организационных моделей. END TAG

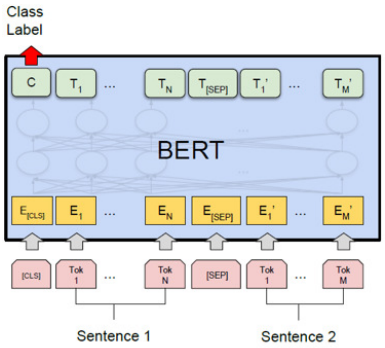
Затекст может выбираться из словаря фраз или свободно генерироваться по контексту, может иметь один или несколько тегов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

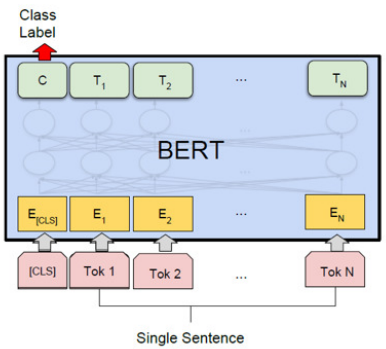
Унификация моделей разметки

Большие пред-обученные модели языка (трансформеры)

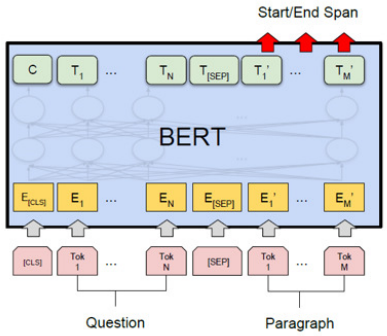
- обучены по терабайтам текстов, «они видели в языке всё»
- способны выделять и классифицировать фрагменты текста
- способны генерировать связный текст
- *мультиязычны*: обучаются на десятках языков
- *мультизадачны*: для каждой новой задачи NLP/NLU достаточно пред-обученной модели + дообучения на относительно небольшой размеченной выборке



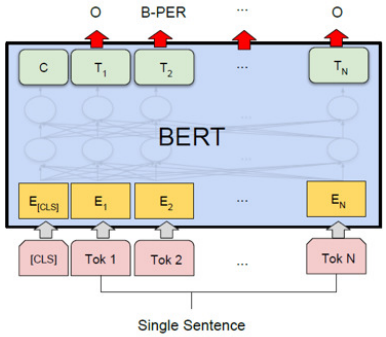
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Унификация оценивания моделей разметки

- В основе методики — сравнение пар разметок текста: «алгоритм – эксперт», «эксперт-1 – эксперт-2», путём оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок $\mathrm{Con}_k(A, B)$
- Вводится их средневзвешенная согласованность $\mathrm{Con}(A, B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по размеченной выборке согласованность $\mathrm{Con}(A, E)$ разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по размеченной выборке согласованность $\mathrm{Con}(E_1, E_2)$ разметок двух экспертов, E_1 и E_2
- ОТАР = СТАР / СТЭР, если больше 100%, то алгоритм не хуже экспертов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Выводы по части 2 (из 3)

Некоторые направления исследований в области языкового ИИ:

- модели вставки точных цитат и ссылок на источники
- модели выявления поляризации и сопоставления мнений
- модели взаимодействия с внешними решателями задач
- модели культурного контекста: коммуникативных норм, мировоззрений, социокультурных и ценностных кодов
- модели разметки для формализации гуманитарных знаний, генерации объяснений и количественной аналитики
- модели детекции искусственно сгенерированного текста
- модели детекции фейков, манипуляций, пропаганды для противодействия угрозам в информационном поле

J.Togelius, G.N.Yannakakis. Choose your weapon: survival strategies for depressed AI academics. April 14 2023.

Тематическое моделирование: «о чём все эти тексты?»

Дано:

- коллекция текстовых документов D, словарь W
- n_{dw} — частота терма $w \in W$ в документе $d \in D$

Найти:

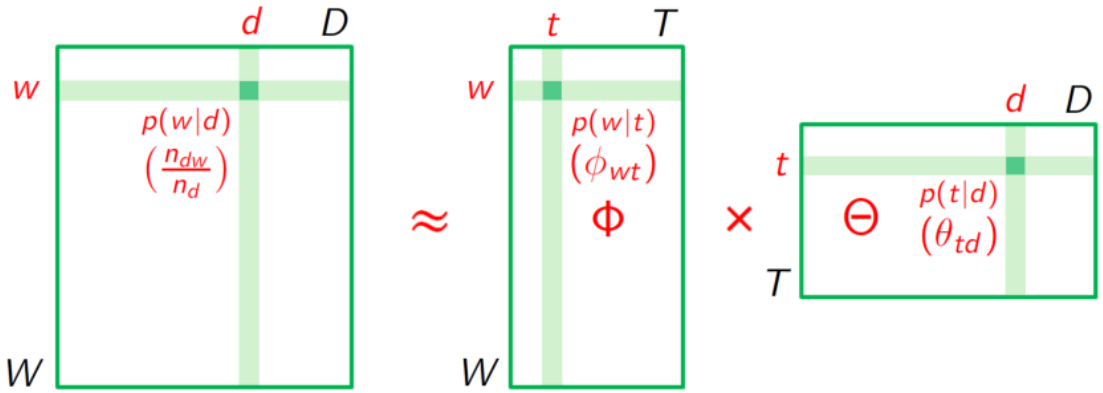
- T — множество тем, составляющих коллекцию D
- $p(w|t) = \varphi_{wt}$ — вероятности слов w в каждой теме t
- $p(t|d) = \theta_{td}$ — вероятности тем t в каждом документе d
- $p(w|d) = \sum\limits_{t \in T} \varphi_{wt}\theta_{td}$ — тематическую языковую модель

Критерий: правдоподобие предсказания слов w в документах d с дополнительными критериями-регуляризаторами $R_i(\Phi, \Theta)$:

$$\sum_{d \in D} \sum_{w \in d} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

Три интерпретации задачи тематического моделирования

1. **Мягкая кластеризация** документов по кластерам-темам
2. Низкоранговое стохастическое **матричное разложение:**



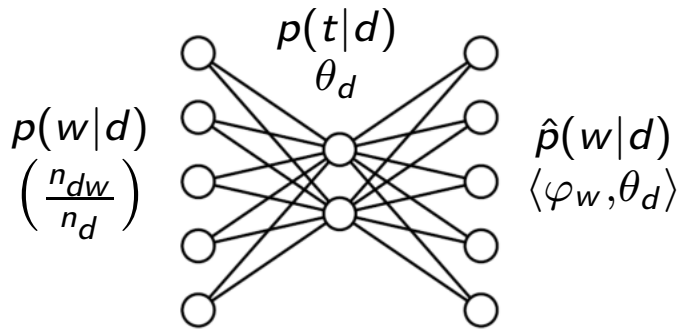
3. **Автокодировщик** документов в тематические эмбединги:

кодировщик $f_\Phi : \frac{n_{dw}}{n_d} \to \theta_d$

декодировщик $g_\Phi : \theta_d \to \Phi\theta_d$

задача реконструкции:

$$\sum_d \mathrm{KL}\left(\frac{n_{dw}}{n_d} \,\middle\|\, \langle \varphi_w, \theta_d \rangle\right) \to \min_{\Phi,\Theta}$$



ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\varphi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \varphi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}$$

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Байесовская и классическая регуляризация

X — данные, текстовая коллекция, $\Omega = (\Phi, \Theta)$ — параметры

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (громоздкий, приближённый) ради точечной оценки Ω:

$$\text{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\text{Prior}(\Omega|\gamma)$$

$$\Omega := \arg\max_{\Omega} \text{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg\max_{\Omega} \big(\ln p(X|\Omega) + \ln \text{Prior}(\Omega|\gamma)\big)$$

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg\max_{\Omega} \Big(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Big)$$

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайновый параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и метрик качества

Сообщество:

- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

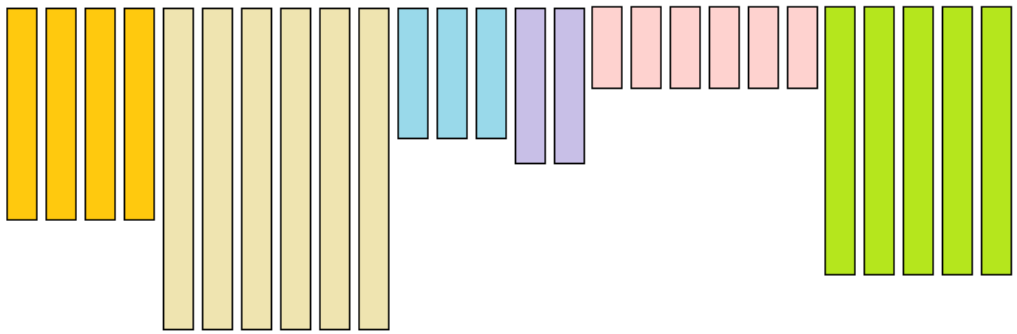
Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \underset{i \in I}{\text{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω.
Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \dfrac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Векторы $\omega_j = 0$ отбрасываются как вырожденные решения
- Итерации похожи на градиентную оптимизацию:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}},$$

но учитывают ограничения и не требуют подбора шага η

EM-алгоритм для ARTM без матрицы Θ

Максимизация логарифма правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td}(\Phi) + R\big(\Phi, \Theta(\Phi)\big) \to \max_{\Phi}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}};$$

$$p'_{tdw} = p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \varphi_{wt} \frac{\partial \theta_{sd}}{\partial \varphi_{wt}};$$

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(\sum_{d \in D} n_{dw} p'_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right).$$

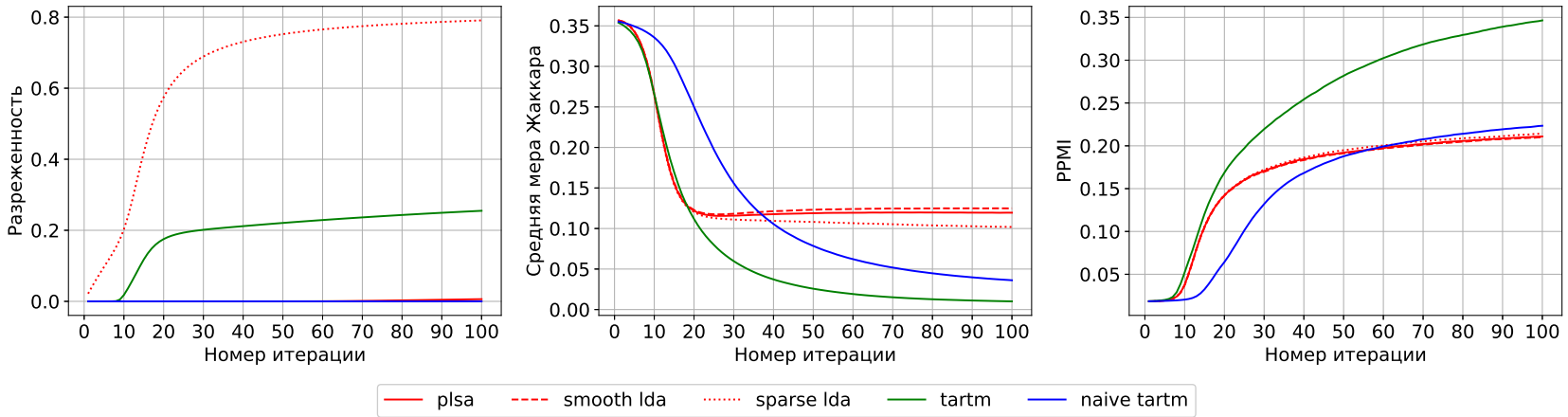
И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КиМ, 2020.

Проверка модифицированного EM-алгоритма

$\theta_{td}(\Phi) = \frac{1}{n_d}\sum_{i=1}^{n_d} \underset{t\in T}{\text{norm}}(\varphi_{w_i t} n_t)$ — линейная тематизация

Эксперимент на коллекции NIPS, $|T| = 50$, модели:

- TARTM (Θless ARTM) — модифицированный EM-алгоритм
- naive TARTM — одна итерация обычного EM-алгоритма



- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность и когерентность тем

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

Однопроходная линейная тематизация + локализация E-шага

$w_1, \ldots, w_n$ — сквозная нумерация термов во всей коллекции
C_i — локальный контекст (окружение) терма w_i
$\alpha(u|i)$ — распределение важности термов $u \in C_i$ для терма w_i

- не нужна гипотеза «мешка слов»
- не нужно разбиение на документы

EM-алгоритм: метод простой итерации для системы уравнений

$$\varphi'_{tw} = \underset{t \in T}{\text{norm}}\left(\varphi_{wt} n_t\right);$$

$$p_{ti} = \underset{t \in T}{\text{norm}}\left(\varphi_{w_i t} \sum_{u \in C_i} \varphi'_{tu} \alpha(u|i)\right); \qquad n_t = \sum_{i=1}^{n} p_{ti};$$

$$\varphi_{wt} = \underset{w \in W}{\text{norm}}\left(\sum_{i=1}^{n} [w_i = w] p_{ti} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right).$$

Сравнение локализованного E-шага с моделью self-attention

Тематический вектор локального контекста на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \underset{t\in T}{\text{norm}}\left(\varphi_{w_i t}\theta_{ti}\right) = \underset{t\in T}{\text{norm}}\left(\sum_{u\in C_i} \varphi'_{tu}\varphi_{w_i t}\alpha(u|i)\right)$$

Вектор контекста (эмбединг) на выходе модели внимания:

$$y_i = \sum_{u\in C_i} Vx_u\,\alpha(u|i) = \sum_{u\in C_i} Vx_u\,\underset{u\in C_i}{\text{SoftMax}}\langle Qx_i, Kx_u\rangle.$$

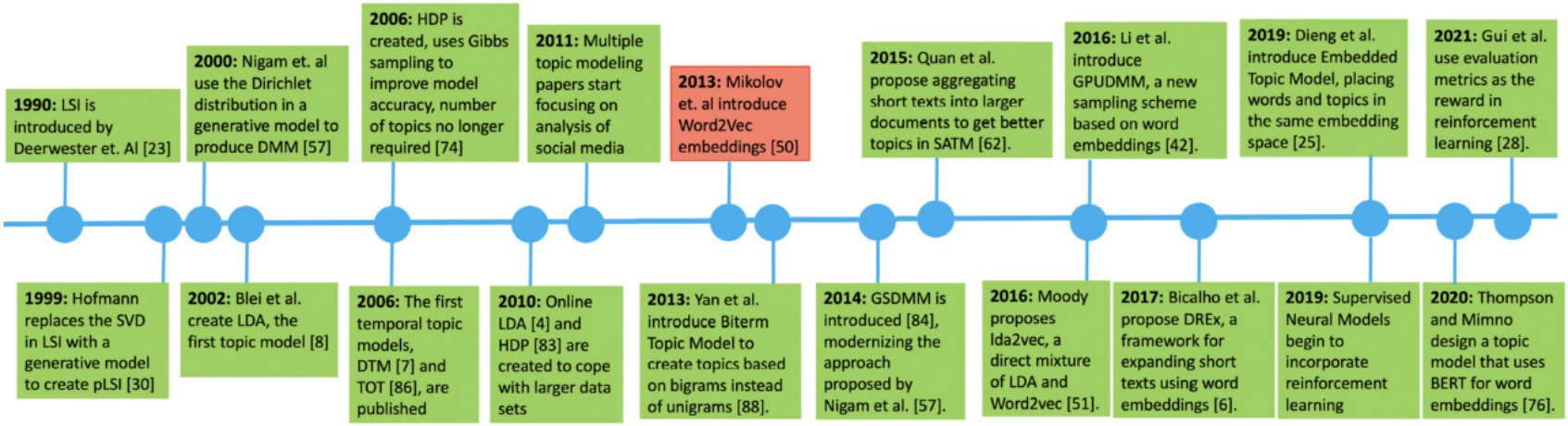
Сходство:

- вектор терма w_i трансформируется в вектор его контекста
- путём усреднения векторов u из контекста терма w_i,
- наиболее (семантически) схожих с вектором терма w_i.

Отличия:

- адамарово умножение вектора φ'_{tu} на вектор-фильтр $\varphi_{w_i t}$;
- нет обучаемых параметров Q, K, V как у модели внимания;
- проецирование итогового вектора на единичный симплекс.

Эволюция тематического моделирования



Neural Topic Models — поток публикаций начиная с 2016

Как «объединить лучшее от двух миров»?

- **Neural:** качество, универсальность, генеративность
- **Topic:** скорость, интерпретируемость, простота

Что объединяет: векторизация, оптимизация, регуляризация, гомогенизация, локализация (контекст и внимание)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Выводы по части 3 (из 3)

Некоторые направления исследований в Topic Modeling

- **Гипотеза:** нормировка векторов в нейронных сетях может привести к появлению свойств интерпретируемости
- Есть успешная первая реализация ARTM на pyTorch, **неожиданно:** на GPU время вычислений не зависит от $|T|$
- От локализации E-шага к тематическим моделям внимания
- Упрощение моделей внимания и архитектур трансформера

Vorontsov K. V. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization. 2023 (принято к публикации)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022

He Zhao, Dinh Phung , Viet Huynh , Yuan Jin , Lan Du , Wray Buntine. Topic Modelling Meets Deep Neural Networks: A Survey. 2021

Liang Yang et al. Graph Attention Topic Modeling Network. 2020

Tian Tian et al. Attention-based Autoencoder Topic Model for Short Texts. 2019

Shuangyin Li et al. Recurrent Attentional Topic Model. 2017