AIRI

Облегчение моделей by design

Елена Тутубалина

Научный сотрудник НИУ ВШЭ,
исполнительный директор по исследованию данных, Sber AI

Александр Петюшко

Директор ключевых исследовательских программ, AIRI

AGENDA

01 Machine Reading Comprehension

02 Retrieval to the rescue

03 Entity Linking

01

Machine Reading Comprehension

Machine Reading Comprehension (MRC) as Explainability *by initio*

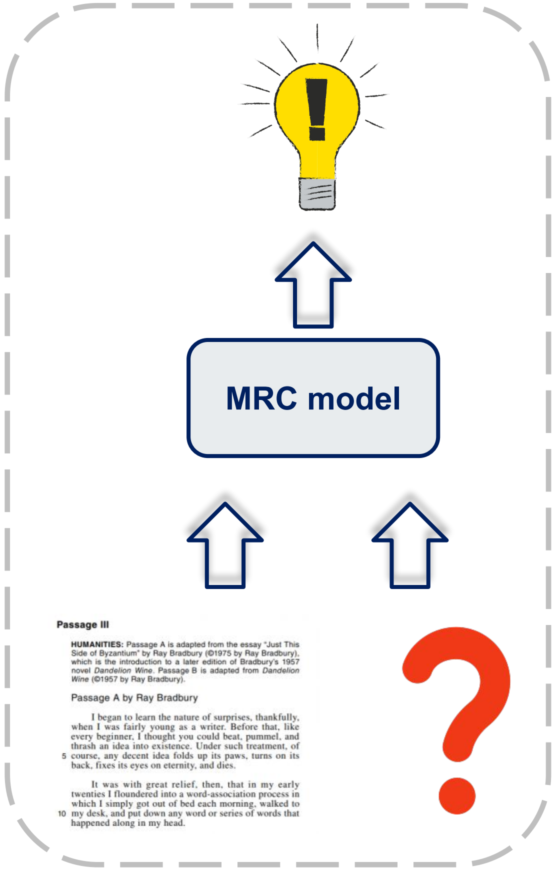
- Question Answering (QA): standard NLP task

- Now most of the best QA-systems are **generation**-based:
  - Means that only large (or even HUGE) **decoder** is used
  - All the information needed to answer the question is stored **inside decoder weights**
  - But the output is **unexplainable**: the model just knows (or not!) the answer

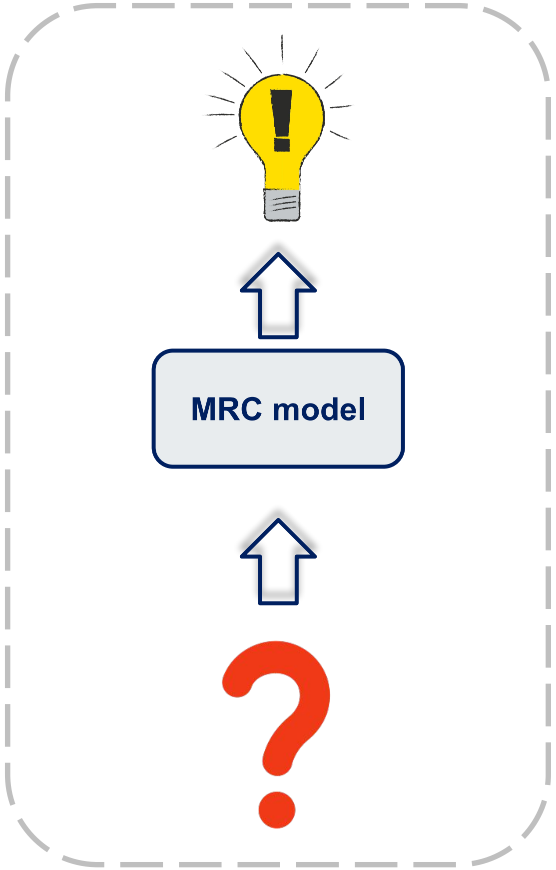
- What we'd like to have: the explainability **WHY** the system provides this answer

- In terms of MRC it means that the system can provide the **relevant** text **passage** (or passages), **containing** the correct answer
  - And the **human** can **understand** whether the system was right about it's guessing
  - At the same time, it can lead to **decreasing** the model **size** (usage of a number of small models is still more efficient than one huge decoder)

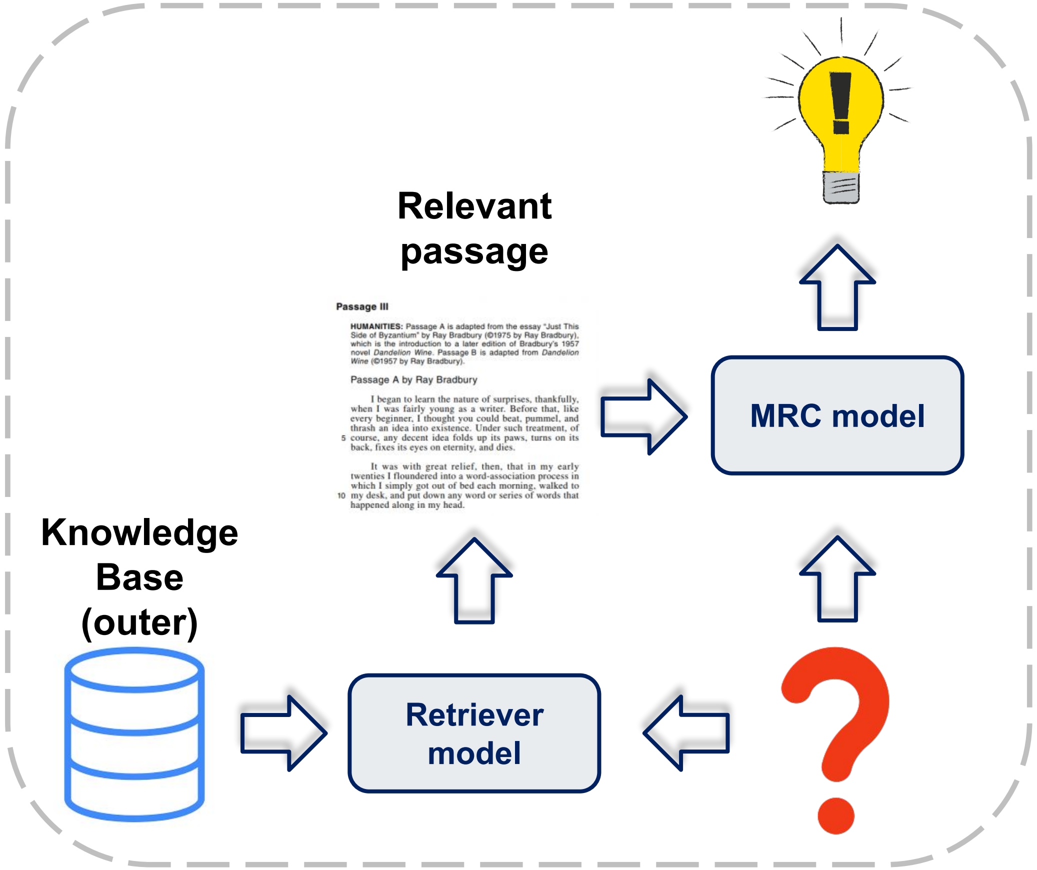
Machine Reading Comprehension: common paradigms



Extraction of knowledge from relevant passage
Not possible in real-world



Generation of knowledge[1,2]
Not scalable, all information is stored inside MRC model weights (like T5/GPT-3)



2-stage: <u>first</u> to <u>retrieve</u> the relevant model from outer text corpus, <u>then</u> <u>extract</u> knowledge from this passage
Realistic, explainable and scalable approach

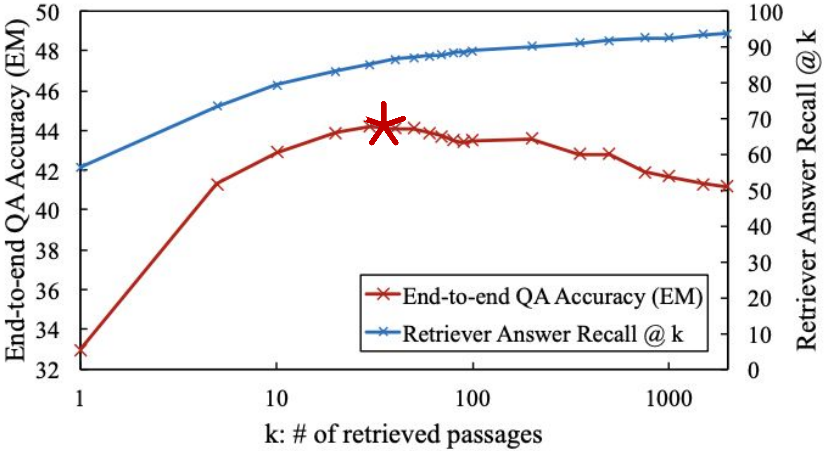
[1] Roberts, Adam, Colin Raffel, and Noam Shazeer. "How Much Knowledge Can You Pack Into the Parameters of a Language Model?."
[2] Brown, Tom B., et al. "Language models are few-shot learners."

02

Retrieval to the rescue!

Retriever ≠ Reader[1]

(a) End-to-end QA accuracy (Exact Match, y-axis on the left) of DPR reader and the retrieval recall rate (y-axis on the right) of DPR retriever.



$$p_\eta(z|x) \propto \exp\left(\mathrm{d}(z)^\top \mathrm{q}(x)\right)$$

$$\mathrm{d}(z) = \mathrm{BERT}_d(z), \quad \mathrm{q}(x) = \mathrm{BERT}_q(x)$$

BERT as a Retriever (DPR)

Main idea:

- Retriever **is not approx.** of Reader: having more data helps a little for the Reader, but then drops quickly
- **Retriever** is a sort of **<u>representational bottleneck</u>**
- Can improve **Retriever** by KD from Reader: helps significantly for retrieval, but not so much for MRC
  - **RDR**: Reader-distilled Retriever
- KD by aligning similarities doc <> query

Retriever improvement after KD

Dataset	NQ-dev				NQ-test				TriviaQA-test			
Top-k	1	20	50	100	1	20	50	100	1	20	50	100
DPR-Single	44.2‡	76.9‡	81.3‡	84.2	46.3	78.4†	84.1	85.4†	54.4	79.4†	82.9	85.0†
↳ w/ RDR	**54.1** (+9.9)	**80.7** (+3.8)	**84.1** (+2.8)	**85.8** (+1.6)	**54.2** (+7.9)	**82.8** (+4.4)	**86.3** (+2.2)	**88.2** (+2.8)	**62.5** (+8.1)	**82.5** (+3.1)	**85.7** (+2.8)	**87.3** (+2.3)
SOTA	51.7‡	79.2‡	83.0‡	-	-	79.4†	-	86.0†	-	79.9†	-	85.0†

Reader improvement after KD

Dataset	NQ-test			TriviaQA-test		
	Top-1	Reported		Top-1	Reported	
	EM	EM	Top-k	EM	EM	Top-k
DPR-Single	32.3	41.5	50	44.5	56.8	50
↳ w/ RDR	37.3 (+5.0)	42.1 (+0.6)	10	49.1 (+4.6)	57.0 (+0.2)	50
RAG-Token	39.4	44.1	15	-	55.2	-
↳ w/ RDR	40.9 (+1.5)	44.5 (+0.4)	15	-	-	-

[1] Yang, Sohee, and Minjoon Seo. "Is Retriever Merely an Approximator of Reader?." (*NAVER Corp*)

Fusion-in-Decoder (FiD)[1]: RB model for MRC

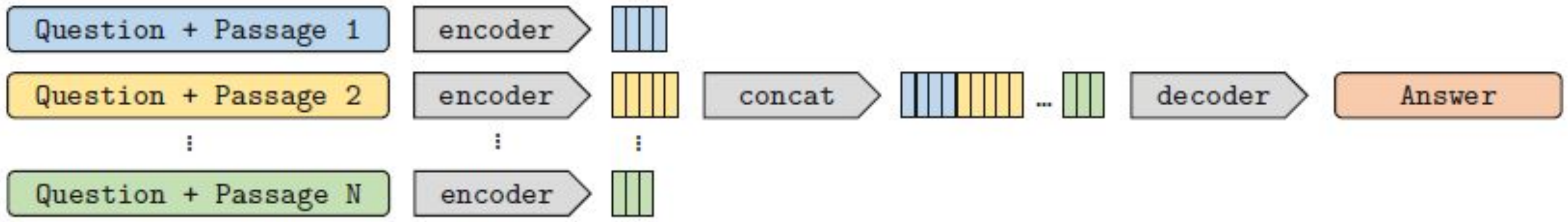
FiD
=
usual retriever
+
generator as reader
+
reading answer from N passages

Main idea:

- **Retriever:** DPR (BERT-doc + BERT-query)
- **Reader** is **seq2seq T5**, having **query + retrieved doc** as an **input**
  - added special tokens - `question:`, `title:` and `context:` before the question, title and text of each passage
- **Fusion-in-Decoder:** output based on **N > 1 passages**

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right)$$
$$\mathbf{d}(z) = \text{BERT}_d(z), \quad \mathbf{q}(x) = \text{BERT}_q(x)$$

BERT as a Retriever (DPR)



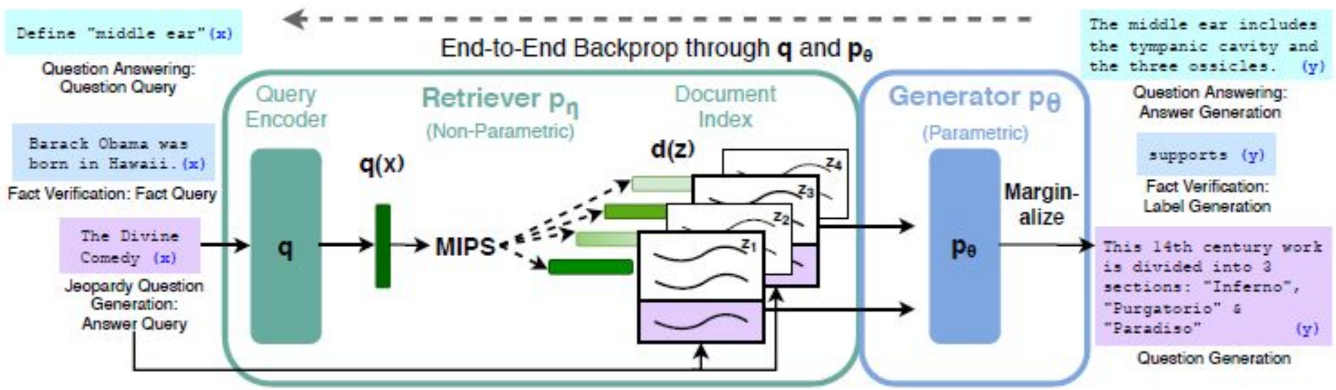
[1] Izacard, Gautier, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." (*Facebook*)

Retrieval-Augmented Generation (RAG)[1]: RB model for MRC

RAG
=
usual retriever
+
generator as reader

Main idea:

- **End-to-end backprop** through **retriever AND reader**
- **Retriever** is initialized from **DPR**[2] approach
- **Reader** is **seq2seq BART**, having **query + retrieved doc** as an **input**
- **Generator** can provide the output based on **1 passage** (Sequence-based) **or k > 1 passages** (Token-based)
- **Better** than **BERT-based reader**, but **more heavy** (400M vs 110M)



Seq2seq generator (BART)
As a Reader

1 passage:

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z\in\text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

k passages:

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z\in\text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z,y_{1:i-1})$$

[1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." (*Facebook*)
[2] Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." (*Facebook*) // ColBERT-like

03

Entity Linking

Biomedical Entity Linking

NIH U.S. National Library of Medicine

ClinicalTrials.gov

Condition or disease	Intervention/treatment	Phase
Squamous Cell Carcinoma of Lung	Drug: Icotinib	Phase 2

Condition or disease	Intervention/treatment	Phase
Non-Squamous Non-Small Cell Lung Cancer	Drug: Erlotinib	Phase 2

Condition or disease	Intervention/treatment	Phase
NSCLC Non-small Cell Lung Cancer	Drug: MEDI4736 (anti-PD-L1)	Phase 2

Condition or disease	Intervention/treatment	Phase
Non-Small Cell Lung Cancer, Ovarian Cancer	Drug: DNIB0600A	Phase 1

Carcinoma, Non-Small-Cell Lung MeSH Descriptor Data 2021

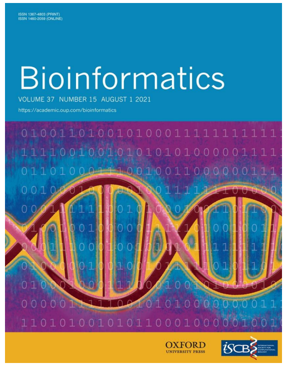
Details Qualifiers MeSH Tree Structures Concepts

MeSH Heading Carcinoma, Non-Small-Cell Lung
Tree Number(s) C04.588.894.797.520.109.220.249
C08.381.540.140.500
C08.785.520.100.220.500
Unique ID D002289
RDF Unique Identifier http://id.nlm.nih.gov/mesh/D002289
Annotation coordinate IM with LUNG NEOPLASMS (IM); CARCINOMA, LARGE CELL and SMALL CELL LUNG CARCINOMA are also available
Scope Note A heterogeneous aggregate of at least three distinct histological types of lung cancer, including SQUAMOUS CELL CARCINOMA; ADENOCARCINOMA; and LARGE CELL CARCINOMA. They are dealt with collectively because of their shared treatment strategy.

Ovarian Neoplasms MeSH Descriptor Data 2021

Details Qualifiers MeSH Tree Structures Concepts

MeSH Heading Ovarian Neoplasms
Tree Number(s) C04.588.322.455
C13.351.500.056.630.705
C13.351.937.418.685
C19.344.410
C19.391.630.705
Unique ID D010051
RDF Unique Identifier http://id.nlm.nih.gov/mesh/D010051
Annotation coordinate IM with histologic type of neoplasm (IM)
Scope Note Tumors or cancer of the OVARY. These neoplasms can be benign or malignant. They are classified according to the tissue of origin, such as the surface EPITHELIUM, the stromal endocrine cells, and the totipotent GERM CELLS.



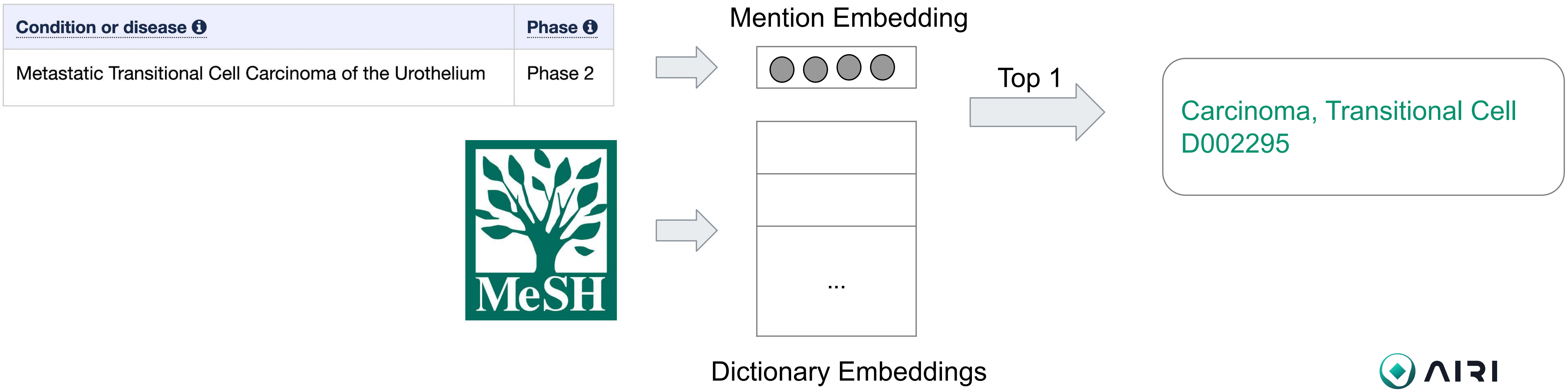
Our approach
DILBERT

Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer

Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, Elena Tutubalina

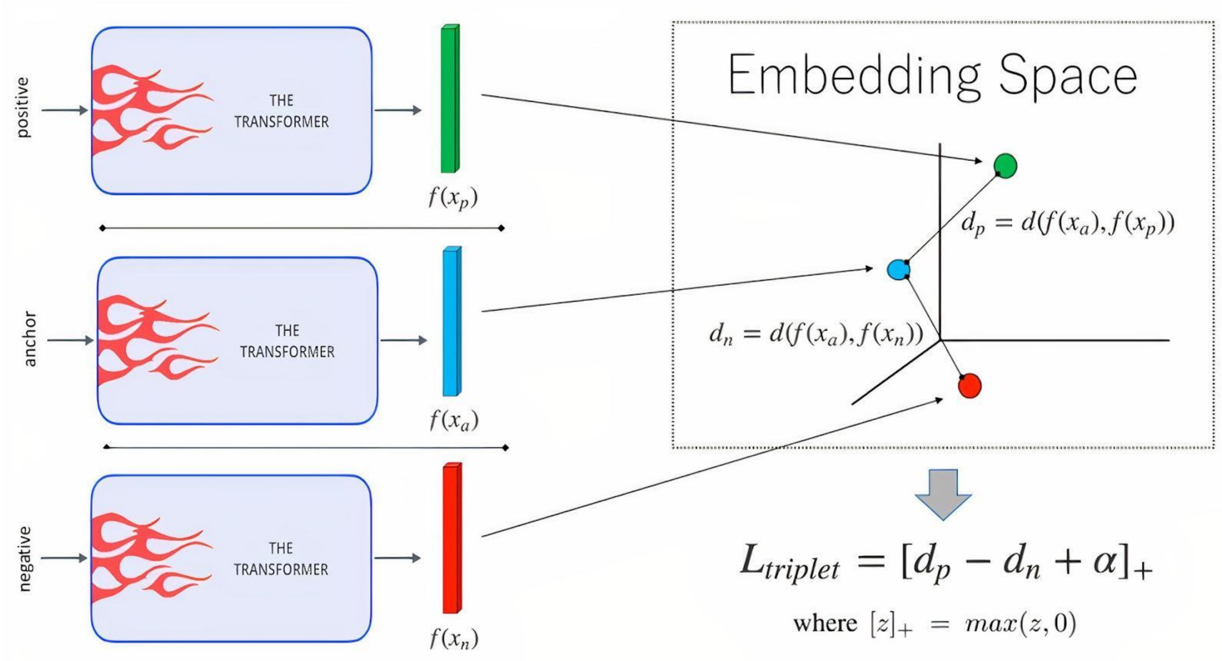
DILBERT - Design

- Most of the best biomedical entity linking systems:
  - are trained & evaluated in the single-terminology setting
  - use classification type losses and online processing (a.k.a. readers)
- We focus on **cross-terminology** mapping of entity mentions to a given lexicon **without additional re-training**
- Fast, **real-time inference** -- all concept names from a terminology are cached



DILBERT - Training

- We use triplets of free-form entity mention, positive and negative concept names



Disease mention

Condition or disease	Phase
NSCLC Non-small Cell Lung Cancer	Phase 2

Positive concept names

Carcinoma, Non-Small-Cell Lung
Non-Small Cell Lung Cancer
Non-Small Cell Lung Carcinoma

The rest of the MeSH dictionary for negative sampling

Carcinoma, Bronchogenic
Lung Neoplasms
Cancer of the Lung
Rhinitis
...

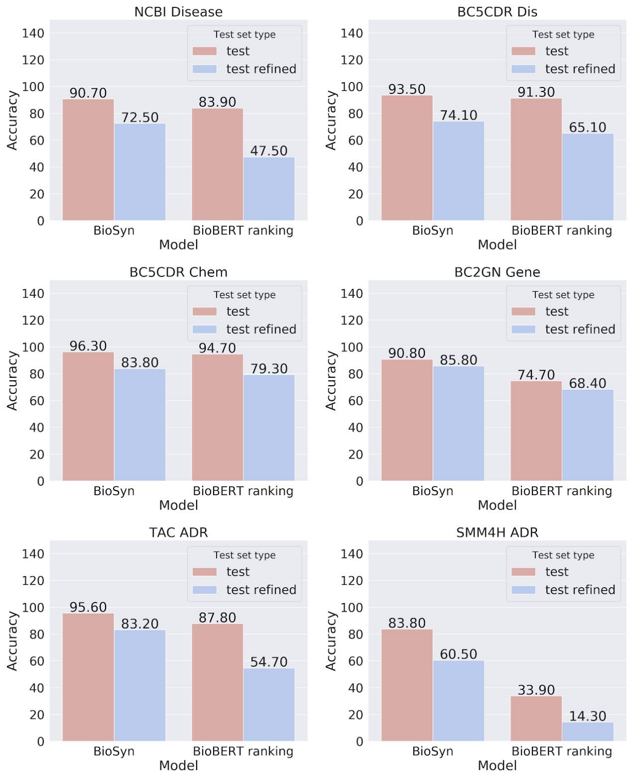
Let's remove bias!

COLING 2020

Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models

Elena Tutubalina, Artur Kadurin, Zulfat Miftahutdinov

- Evaluation of benchmarks: BioCreative V CDR, BioCreative II GN, NCBI Disease, and TAC 2017 ADR
- App. 80% entity mentions in the test set are textual duplicates of other entities presented in the test set or train+dev sets
- Divergence in performance between these the original and ***refined*** test sets (app. 15%)
- Propose *cross-terminology* evaluation

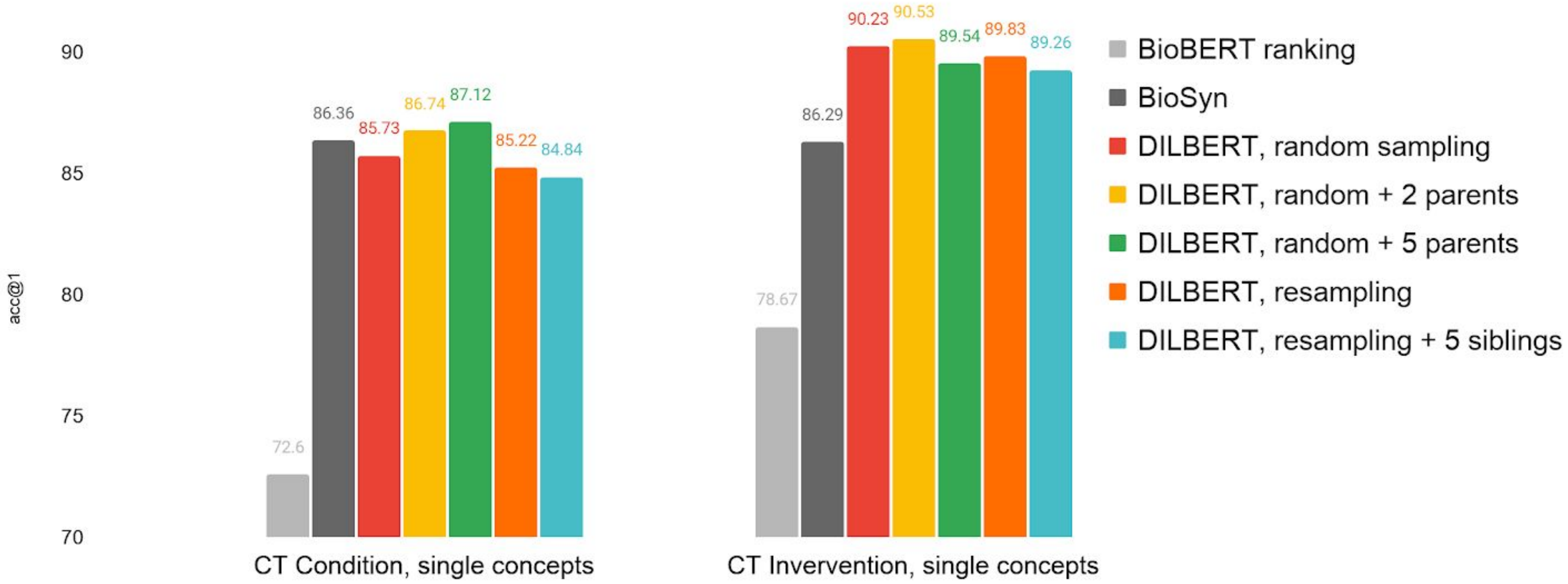


https://www.aclweb.org/anthology/2020.coling-main.588.pdf

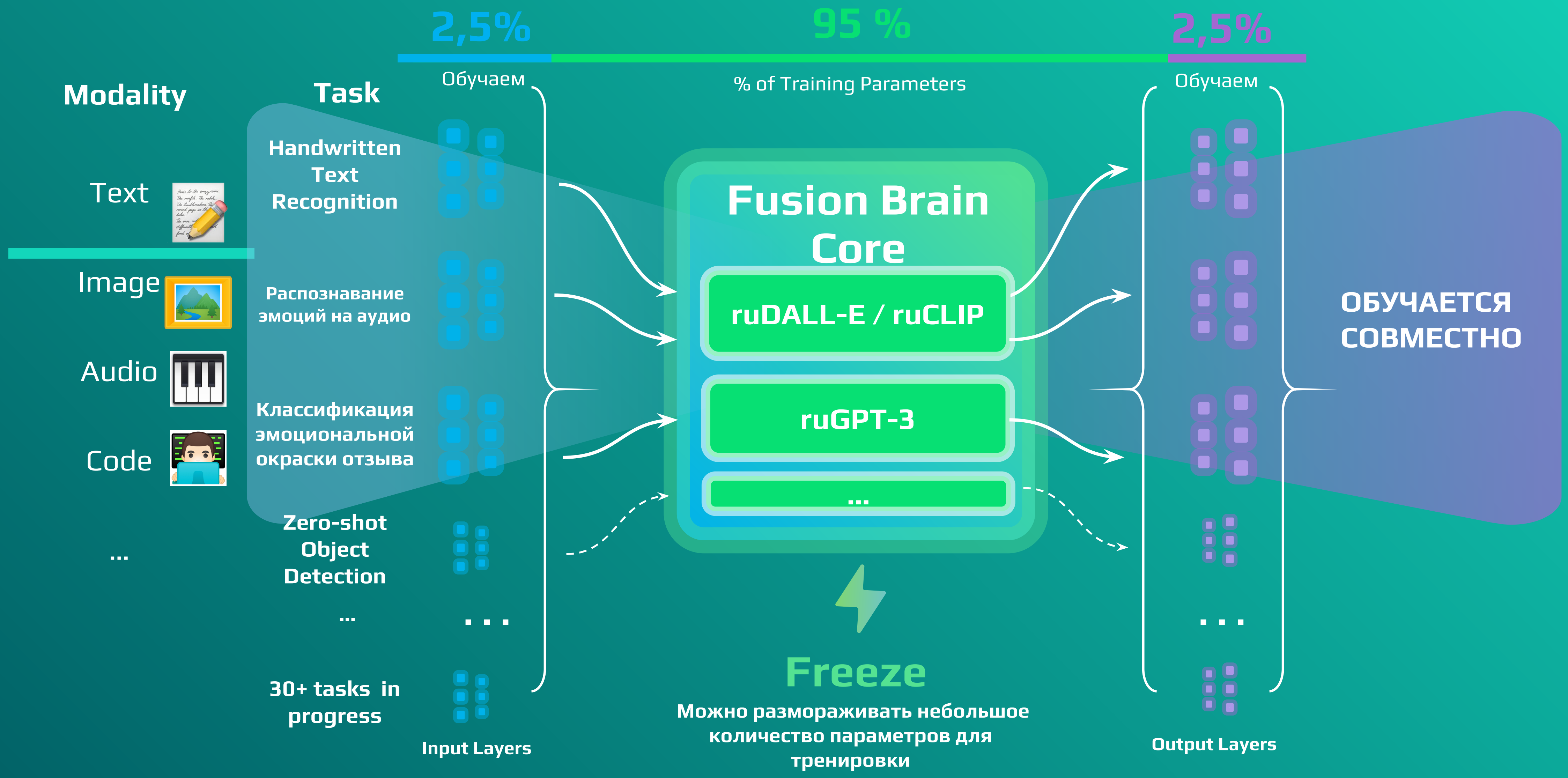
Russian Science Foundation project no. 18-11-00284

AIRI

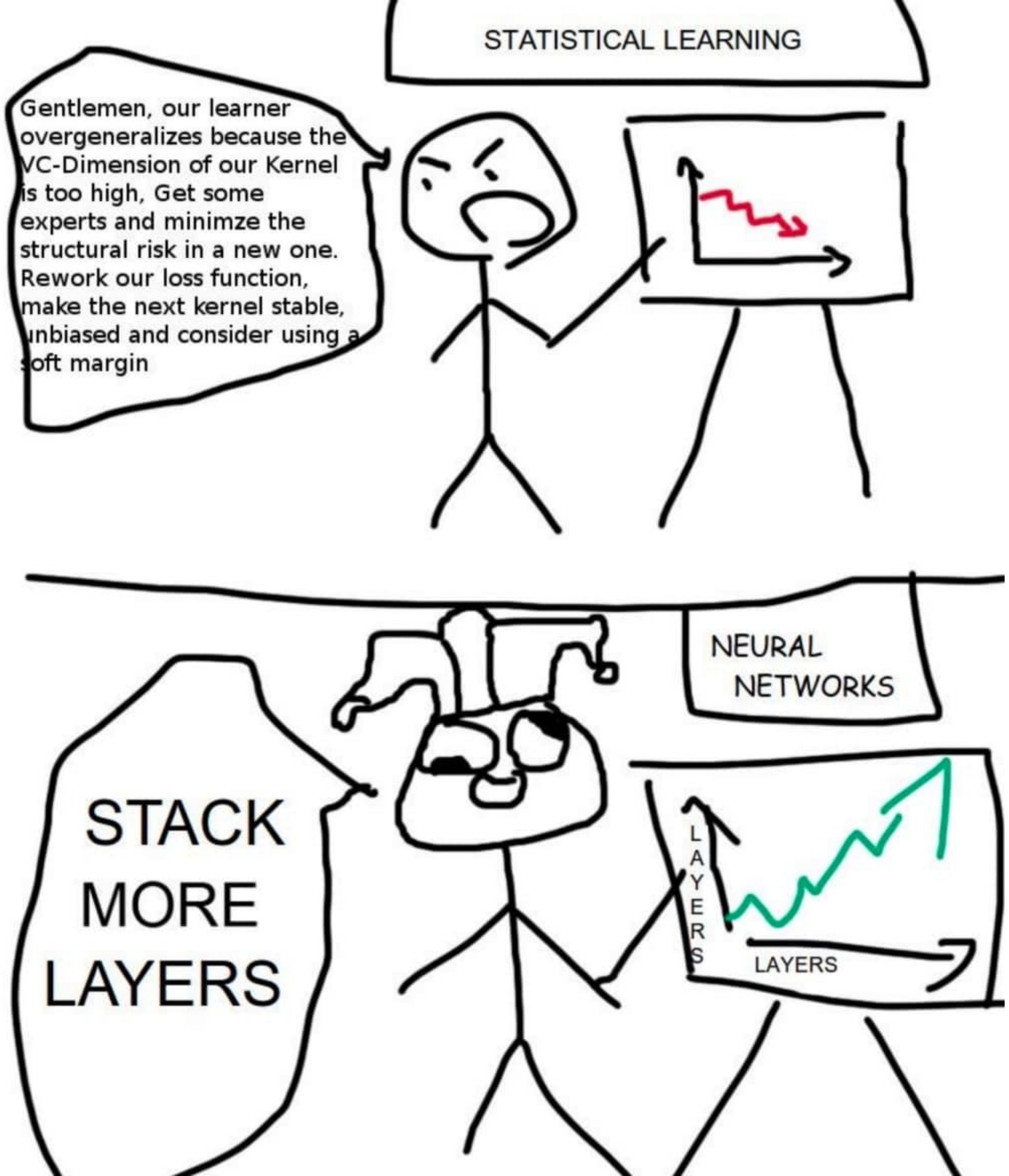
Experiments



Fusion Brain: Effective Multi-modal Multi-task model



https://github.com/sberbank-ai/fusion_brain_aij2021


STATISTICAL LEARNING
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimze the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin
NEURAL NETWORKS
STACK MORE LAYERS
LAYERS
LAYERS