Many heads but one brain: FusionBrain – a single multimodal multitask architecture and a competition

D.D. Bakshandaeva[1,4]*, D.V. Dimitrov*[1,2,6]*, V.S. Arkhipkin*[1]*, A.V. Shonenkov*[2]*,*
M.S. Potanin[2]*, D.K. Karachev*[2]*, A.V. Kuznetsov*[1,2,3]*, A.D. Voronov*[2]*, A.A. Petiushko*[2]*,*
V.F. Davydova[1]*, E.V. Tutubalina*[1,2,5]
[1]*Sber AI, 121170, Moscow, Russia, Kutuzovsky prospekt, 32, building 2;*
[2] *Artificial Intelligence Research Institute, 105064, Moscow, Russia, Nizhniy Susalnyy pereulok, 5;*
[3]*Samara National Research University, 443086, Samara, Russia, Moskovskoye Shosse, 34;*
[4]*University of Helsinki, 00014, Helsinki, Finland, Yliopistonkatu, 3;*
[5]*National Research University Higher School of Economics, 109028, Moscow, Russia, Pokrovsky Bulvar, 11;*
[6]*Moscow State University, 119991, Moscow, Russia, Kolmogorova, 1*

Abstract

Supporting the current trend in the AI community, we present the AI Journey 2021 Challenge called FusionBrain, the first competition which is targeted to make a universal architecture which could process different modalities (in this case, images, texts, and code) and solve multiple tasks for vision and language. The FusionBrain Challenge combines the following specific tasks: Code2code Translation, Handwritten Text recognition, Zero-shot Object Detection, and Visual Question Answering. We have created datasets for each task to test the participants' submissions on it. Moreover, we have collected and made publicly available a new handwritten dataset in both English and Russian, which consists of 94,128 pairs of images and texts. We also propose a multimodal and multitask architecture – a baseline solution, in the centre of which is a frozen foundation model and which has been trained in Fusion mode along with Single-task mode. The proposed Fusion approach proves to be competitive and more energy-efficient compared to the task-specific one.

Keywords: multimodality, multitask, bilinguality, foundation models, FusionBrain challenge.

Citation: Bakshandaeva D, Dimitrov D, Arkhipkin V, Shonenkov A, Potanin M, Karachev D, Kuznetsov A, Voronov A, Petiushko A, Davydova V, Tutubalina E. Many heads but one brain: FusionBrain – a single multimodal multitask architecture and a competition. Computer Optics 2023; 47(1): 185-195. DOI: 10.18287/ 2412-6179-CO-1220.

Introduction

A significant part of the information perceived by a person and required for making even the simplest everyday decisions is presented in multiple modalities, that is, with the help of different types of "input information", requiring the use of various senses and types of knowledge. Visual information requires visual perception, processing natural language texts presupposes the knowledge of the language, auditory information implies the perception and analysis of sound, and so on. Each of these modalities is handled by separate, sometimes overlapping areas of machine learning and artificial intelligence: computer vision, natural language processing, speech processing, video processing, etc.

However, a successful solution to emerging problems often can't be obtained by analyzing data coming from only one modality, just as it is not always sufficient for a human being to use only sight or only hearing to make a rational decision [1]. In such cases, information required to solve such problems can be divided into several "input types", called data modalities, all of which should be taken into consideration to make successful decisions.

Multitask learning has a long history mostly in the natural language processing domain. One of the possible reasons is that by having the correct representation and thus "understanding" of text passage, one can solve many downstream tasks: sentiment analysis, question answering, language translation etc. One of the most widely used approaches here is to have the lower (encoding) layers shared for all tasks, while having the upper layers (also called "heads") task-specific and learned separately [2].

It is only recently that scientists have proposed to combine multimodality and multitask in one model, taking the joint approach: using different encoders for different modalities, then combining different types of information during middle processing, and completing the process with task-specific heads – e.g. the UniT [3] approach, where visual and textual modalities are used, and 7 tasks of computer vision (e.g. object detection), text processing (e.g. sentiment analysis) and vision-and-language (e.g. visual question answering) fields are solved.

The problem of training large multimodal and multitask models can be separated into 2 subtasks: 1) How to combine modalities, and 2) How to combine tasks.

As for the first question, the current state-of-the-art research in the multimodal processing is mostly focusing on the questions of the stage at which modalities should be fused ("early", "middle" or "late" fusion) and the ways to implement this fusion (through iterative processing or by a modality bottleneck) [4, 5, 6, 7]. One of the most interesting approaches for modality fusion are Perceiver [8] and Perceiver IO [9], where the modality-specific information serves as the key-value for iterative cross-attention and is later processed by GPT-2-like [10] transformer. Cross-attention blocks are also used in Flamingo [11] to incorporate the information from the pretrained visual encoder into the frozen language model (Chinchilla [12]), thus allowing to process multimodal (visual and textual) inputs.

Another interesting and promising example of sharing the modality information is the so-called multimodal bottleneck transformer (MBT) [13], where the fusion of the modalities is done: a) closely to the top of the transformer layers; b) only through a very small number B of multimodal neurons (in the work $B=4$ is used) implementing the cross-modality sharing only through a small bottleneck, which proves to be very efficient. Finally, the incorporation of different modalities (like RGB and OpticalFlow) inside the single model via mutual modality learning can be used [14].

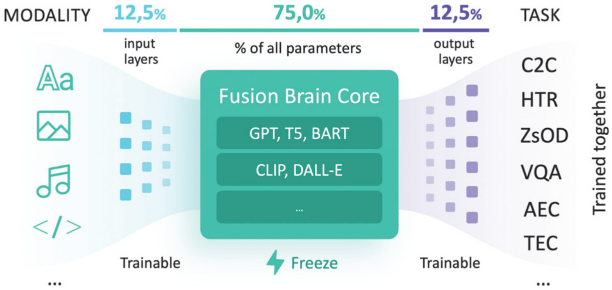


Fig. 1. Concept of the multimodal and multitask architecture FusionBrain. The tasks here are C2C – Code2code Translation, HTR – Handwritten Text Recognition, ZsOD – Zero-shot Object Detection, VQA – Visual Question Answering, AEC – Audio Emotion Classification, and TEC – Text Emotion Classification

The combination of tasks can also be implemented in different ways. An approach similar to the above-mentioned UniT is the so-called frozen pretrain transformer (FPT) technique [15], which is a source of inspiration for our proposed baseline. However, such multitask pipeline, when different tasks/modalities are processed through separate heads, is not the only one. The more interesting approaches use more sophisticated ways of dealing with multiple tasks: for instance, the task-specific adapters [16, 17] between the frozen layers can be used or the fully learnable (trainable) task representation (embedding) can be incorporated and later propagated in a non-trivial way through the major part of the model (see Perceiver IO, HyperGrid [18] or conditionally adapted approach [19]).

As different tasks within the domain can have similar formats, in general-purpose agent Gato [20], a transformer decoder model, it is proposed to use prompt conditioning instead of simple one-hot identifiers in order to disambiguate tasks: while training, for 25 % of the sequences in the batch, a prompt sequence generated by the same agent on the same task, is added. In half of these cases, the sequence is taken from the end of the episode (goal conditioning is obtained), in the other half, it is randomly sampled from the episode. Using prompt conditioning, as the authors note, would be ideal for adjusting to new tasks, if not the model's maximum context length restraint which doesn't allow the agent to get access to the information sufficient to solve the desired problems.

The corresponding research in the field of information retrieval (IR) is also worth mentioning. For now, however, it seems that quite straightforward solutions are used for IR, e.g. the combination of all task-specific datasets for training NLP model for multiple tasks [21], or the processing of multimodal data with the single transformer using the representations obtained by modality-specific encoders as the inputs for the multimodal retrieval [22, 23].

We aim to promote the development of such promising and challenging field as multimodal and multitask research. Our main contributions are the following:

- preparing the data, task statement and leaderboard for the FusionBrain Challenge;
- proposing the specialized as well as the overall metric to evaluate the models;
- creating a simple yet efficient baseline which combines multimodal as well as multitask approach.

1. Tasks

Within the competition we proposed to solve 4 subtasks:

1. Code2code translation (C2C).
2. Handwritten text recognition (HTR).
3. Zero-shot object detection (ZsOD).
4. Visual question answering (VQA).

In order for the model presented by the team/participant to be considered as multitask, it is necessary and sufficient to meet the following criteria:

1. Shared weights should be at least 25 % of all model parameters: if N is the total number of parameters of the models that solve 4 subtasks, and M is the number of common parameters of these models (that is, they are identical both in value and architecturally), then it is necessary that $M/N \geq 0.3$.
2. Common parameters should not be purely nominal – on the contrary, they should be used in a

meaningful way during the prediction of the model and have a beneficial effect on a model's quality.

The participants were provided with an access to GPU resources upon a request. All subtasks which include natural language data are bilingual – contain samples in both English and Russian. In the following subsections, we will discuss each of the subtasks in more detail.

1.1. Subtask 1 – Code2code translation

Among the various problems within ML4Code field, the task of translating code snippets from one programming language (PL) to another was chosen. Even though source code can be attributed to text modality, it is definitely more structured than natural language, thus we would like to distinguish between them. The proposed task not only adds "code modality" to the challenge but also imposes the requirement for the model to be multilingual since it has to understand and generate code in two PLs.

Our C2C task requires a model to translate code snippets from Java to Python. The choice of such a pair of PLs induces extra complexity to the problem since translation between statically- and dynamically-typed languages is more intricate than translation between PLs with the same type checking.

For training we proposed to use a dataset presented in [24]. AVATAR is a parallel corpus that consists of solutions written in Java and Python for 8,506 programming problems collected from competitive programming sites, online platforms, and opensource repositories. We used solutions of 6,807 tasks from AVATAR for train, leaving 1,699 examples for the public part of the test set. The private test dataset was designed as follows: at first, Python snippets with a length corresponding to that of the 90th percentile of AVATAR test set part written in Python (up to 282 tokens obtained after tokenization [25]) were retrieved from CodeNet [26] dataset; these code snippets were translated to Java by three annotators and then cross-checked; at the final stage, Java functions (not longer than 356 tokens, which matches the 90th percentile of the public test requests' lengths) were back-translated to Python and cross-checked as well to ensure that Python snippets generate the same outputs as source functions when given the same inputs. The resulting number of Java-Python pairs is 322.

CodeBLEU [27] is selected as an evaluation metric for this task.

1.2. Subtask 2 – Handwritten text recognition

Handwritten Text Recognition is the task that naturally combines image and text modalities; the model is given an image with a handwritten piece of text in Russian or English and is required to transcribe it into digital text as an output. The dataset for this task was manually collected and annotated; it is composed of examples from school notebooks. The training data consist of 66,599 images of words written in Russian language (participants of the Challenge could use open datasets containing handwritten English text, e.g., IAM Handwriting Database [28]). The public test set includes 14,973 images: 5,973 in English and 9,000 in Russian. The private test part consists of 12,556 images, 5,494 of which are in English and 7,062 – in Russian. In total, our new handwritten dataset contains 82,661 images of Russian words, which makes it the largest Russian handwritten dataset in the world so far. We have also released this dataset [29] for the benefit of the research community.

The evaluation metric for this task is string accuracy - the proportion of cases in which the predicted text (string) coincides with the ground truth transcription.

1.3. Subtask 3 – Zero-shot object detection

ZsOD task sets the following problems to the model: firstly, the model should accurately predict bounding boxes for various objects depicted in the images, given the descriptions of these objects in natural language [30]. In our case, such a common computer vision task as object detection is complicated by the fact that there is no set of predefined classes to choose from – a model is expected to detect classes not present in the training set (i.e. in a zero-shot regime). During inference, a model receives image-query pairs; a query is formatted as a list of textual descriptions (in Russian or English) of objects to detect. The query may contain entities that are absent in the image; a model should predict an empty list as a bounding box for such objects.

The public test dataset is formed from a part of the VisualGenome [31] dataset (1,000 examples); the set of classes in it was hidden from the participants. Region descriptions from VisualGenome are used as positive classes (descriptions are normalized: reduced to lowercase; non-printable characters are removed, etc.; boxes related to the same entity are combined under a single description); negative classes are formed by replacing some objects/attributes in the description with those that are missing in the photo. For example, "a grey chair" is replaced by "a *pink* chair". Also, descriptions of objects belonging to the same domain as the correct classes are used as negative examples: if the photo shows a street, then as negative examples there may be, for instance, descriptions such as "tall green bricks wall", "shingled home in the distance", "food stand in the street" (provided, of course, that the described objects are not in the photo). The images for the private test set were either extracted from YFCC100M dataset [32] or crawled from the Internet. In total, 827 images were attributed with positive (the descriptions of objects which are present in the photo) and negative (the descriptions of missing objects) labels by 10 annotators. The number of positive classes varies from 7 to 10 – the same held true for the negative ones. For a specific image, descriptions can be either in English or in Russian. There can be more than one bounding box for a particular description in the queries, a perfect model should predict all of them.

To assess the quality of the detection model we use an F1-score:

$$F1 = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}.$$

The F1-score is calculated based on Recall and Precision, which, in turn, depend on a set of prediction statistics – true positive (TP), false positive (FP) and false negative (FN):

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative},$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}.$$

In our non-trivial case of multilabel object detection we calculate these statistics as follows:

- FN – for a given label the model has not predicted or predicted not all required bounding boxes;
- TP – a bounding box predicted by the model has IoU-score (intersection-over-union) with at least one of the ground truth bounding boxes for considered label higher than 0.5;
- FP – a predicted bounding box has IoU score less than 0.5 with all ground truth bounding boxes or there is no object of the given label on the image, yet model has predicted boundaries for it instead of returning empty list.

1.4. Subtask 4 – Visual question answering

VQA is a classical multimodal task that requires model to understand a textual question and generate an answer to it based on the corresponding image. The peculiarity of the problem is that the questions are not homogeneous: a correct answer can either consist of several words, or be monosyllabic (a "yes / no" answer) or be a number. It is assumed that only one answer per question is required. As with other tasks, the model should be bilingual in order to perform well, since questions can be expressed in both English and Russian and the answer is expected to be in the same language except when the question concerns the text on the image. For example, when the question is "What is written on the T-shirt?" the answer should be in the same language in which the text is written.

The public test dataset consists of questions in both Russian and English: the Russian-language part is translated examples from the first 10 thousand samples of the validation part of the VQA v2 dataset, the English part – next 10 thousand original samples from the same dataset. The public test set size is 5,446 examples. The private test set was compiled similarly to the one for ZsOD task, except for the nature of annotation: for each image (1,000 images in total), 6 questions in Russian or English and corresponding answers were formulated, resulting in 6,000 samples. The intersection with the private test set for ZsOD task is 724 images.

The evaluation metric for this task is accuracy. Each question has a list of possible correct answers; if the prediction matches at least one of the ground truth answers, it is considered true positive.

2. Baseline

We provide a concept [33] of a single model that is trained on several tasks related to different modalities (visual, audio and text). The concept is inspired by a work [15] that examines the ability of pretrained language models based on the Transformer architecture to form qualitative representations of arbitrary data sequences – thus, generalizing to other modalities with minimal finetuning. The basis of the architecture proposed in the concept is the pretrained GPT-2[10] language model; experiments are carried out with a model which feed-forward layers are frozen.

We build our baseline solution also on top of Frozen Pretrained Transformer. The overall architecture can be seen in Figure 2. The core, the "shared brain" of the whole pipeline is GPT-2 Medium, pretrained on natural language; each type of data for a particular task undergoes its specific transformations in order to match the GPT-2's input format, and also has its specific head to generate predictions in accordance with the task. The input and output layers for each of the subtasks are described below.

It is worth mentioning that one can use any of the so-called foundation model (see, e.g., in-depth report [34]) instead of GPT-2 as FusionBrain Core (see Figure 1). Following the researchers from Stanford University CRFM we define foundation models as models trained on broad data at scale such that they can be adapted to a wide range of downstream tasks. Good examples of such models are BERT [35], BART [36], T5 [37], GPT-3 [38], CLIP [39], DALL-E [40].

A research on Gato [20], which became publicly available in 2022, when FusionBrain Challenge had passed and a baseline model had been released, proves that such approach – converting data of different modalities into flat sequence of tokens and then processing it with a single transformer decoder – has great potential: a model with a single set of weights can solve 450 out of 604 tasks it was trained on, at over a 50% expert score threshold.

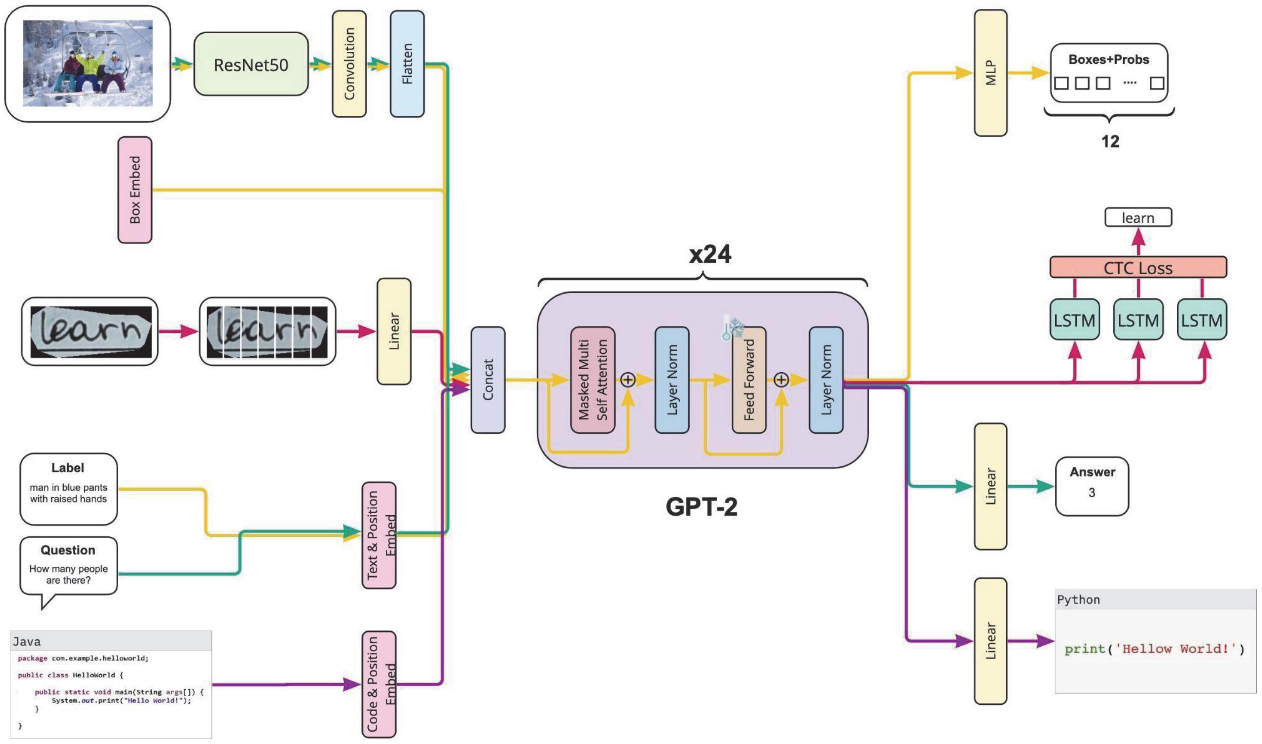
2.1. C2C (code)

As code is similar to natural language (although it is certainly more structured; the problem of choosing the best representation of source code goes beyond the scope of this work), no major transformations are needed in order to prepare the data for processing with GPT-2. The task is solved in decoder-only machine translation manner: during training, the source sequence (code snippet in Java) is concatenated with the target one (in

Python) through the SEP token; the resulting sequence is fed into the GPT-2 with LM head on top in order to minimize the Categorical Cross-Entropy (CCE) loss [41]. When trained, the model auto-regressively generates Python code given Java function.

2.2. HTR (image)

It is somewhat remarkable that images can also be processed using a language model and the proposed method. At first, raw images are subjected to smart resizing with proportions being preserved and empty space being padded; these resized images are then converted into vertical patches with full height and width equal to 8 pixels: $3\times H_0\times W_0\rightarrow 3\times 128\times 512\rightarrow 64\times(128\times 8\times 3)$. Image patch features are extracted with a linear projection layer in order to match the size of the GPT-2 embedding space (1280) before being processed with GPT-2. The transformer outputs are then passed through LSTM and linear layers. The training process is based on the Connectionist Temporal Classification (CTC) loss [42] that shows high performance in handwritten text recognition task [43, 44, 45, 46].



```
package com.example.helloworld;

public class HelloWorld {

    public static void main(String args[]) {
        System.out.print("Hello World!");
    }

}
```

```
print('Hellow World!')
```

Fig. 2. Baseline architecture

2.3. VQA and ZsOD (image+text)

The proposed pipelines for solving VQA and ZsOD tasks are similar. Raw images are resized and processed with a convolutional backbone (three blocks of ResNet-50) [47]; the resulting image feature map is then passed to Conv2D layer with a kernel size equal to 1 and Flatten layer in order to match the size of the embedding space before processing with GPT-2 Medium: $3\times H_0\times W_0\rightarrow 3\times 224\times 224\rightarrow(14\times 14)\times 1024\rightarrow 196\times 1024$. Texts are converted to tokens with the pretrained GPT-2 tokenizer, processed with token and position embeddings. The text format for VQA task is the following: *"Question:"* + *question* + *"Answer:"* + *answer* + *"."*; for ZsOD task: *"Request:"* + *text* + *"."*. The image and text embeddings are concatenated into one sequence: in the case of VQA, text embedding follows image embedding; in the case of ZsOD, it is vice versa.

For VQA, the transformer outputs corresponding to text embeddings are passed through the linear layer in order to get a projection which is consistent with the dimension of the vocabulary. Cross-Entropy loss is used when adjusting model weights during training – and only for the text tokens. The answer is generated auto-regressively.

For ZsOD, 12 trainable box embeddings are introduced and concatenated with the image and text embeddings before being fed into the transformer. The output of GPT-2 is passed through MLP – the resulting dimension is 12×5: for each of 12 boxes, 4 coordinates and a probability score are obtained. The loss function used is similar to the one introduced in [48]: given M ground-truth boxes and N predicted boxes ($N\geq M$), M predicted boxes are chosen so that

$$IoU\left(gt\_boxes\left[:,:4\right], pred\_boxes\left[:,:4\right]\right) \tag{1}$$

$$-pred\_boxes\left[:,-1\right] \tag{2}$$

$$+L1\left(gt\_boxes\left[:,:4\right], pred\_boxes\left[:,:4\right]\right) \tag{3}$$

is minimal.

For the selected boxes, GIoU [49] and L1 losses are minimized; for the probability score, Binary Cross-Entropy Loss is used (1 is assigned to the selected boxes, 0 – to the rest).

3. Experiments

The main goal of our experiments is to compare the metrics of models trained separately for each task and the model trained on all tasks at once (Fusion). We also would like to test the assumption that the combination of similar tasks (in our case, image + text tasks: ZsOD + VQA, and tasks with an image part: HTR + ZsOD + VQA) is the most beneficial for them.

For C2C task, we use AVATAR dataset [24]. While the authors utilize at most 5 accepted solutions for each problem from AtCoder, Code Jam and Codeforces, we raise this number to 7 in order to increase the training dataset. For HTR, samples from IAM Handwritten Database [50] are used. For image-and-text tasks (ZsOD, VQA), we experiment with Visual Genome dataset [31]; for VQA we also add "yes / no" questions from VQA v2 dataset [51]. For testing, we use English-language subsets of the datasets described in 1, in order to compare the results with those produced by state-of-the-art single-task models. The total number of training samples for each task is presented in Table 1.

In Single-task mode, all tasks are trained until the loss reached the plateau, except for ZSoD, since it requires significantly more time for convergence. In Fusion experiments, WeightBalanceSampler is used to avoid unbalanced learning. The sampler weights (see Table 2) are selected based on Single-task training so that in Fusion mode the data for each of the tasks are passed through the model as many times as in Single mode. AdamW optimizer and OneCycleLR scheduler are used for optimization. The following parameters are equal for all experiments (single and fusion tasks): warmup 0.1, pct_start 0.1, max lr 1e-3, start_lr=8e-6, weight decay 1e-2, beta coefficients (0.9,0.999), final_div_factor=1000, 8xA100 80Gb GPUs.

Tab. 1. Number of training samples for different subtasks

	C2C	HTR	ZsOD	VQA
# of samples	92,307	139,917	3,220,243	1,663,852

Tab. 2. Weights of WeightBalanceSampler for different tasks

training setup	C2C	HTR	ZsOD	VQA
ZsOD + VQA	–	–	0.78	0.22
HTR + ZsOD + VQA	–	0.19	0.64	0.17
Fusion	0.04	0.18	0.61	0.17

The results of our experiments are introduced in Table 3. A total score is the sum of scores for four subtasks. Since all tasks are scored from 0 to 1 (the only exception is the CodeBLEU metric: it may take values within the range from 0 to 100 – with a view to normalizing it, the metric is multiplied by 0.01), the final result can range from 0 to 4.

We also measured the performance of state-of-the-art single-task models – PLBART [52], Easter2 [53], MDETR [48] – for each of the subtasks on our private test sets (see Table 4). It should be noted that the vast majority of models (including the state-of-the-art one) solve the VQA task as a classification problem, which is much easier than a generation (the case of our model), but at the same time, such a design of a problem is far from a real application (as in real cases the questions can be very different, and the exhaustive set of answers – "classes" – can't be picked in advance). Although SOTA single-task models show higher scores for each of the subtasks (especially for the VQA task, for the reasons stated earlier, considering the fact that English parts of our public and private test sets were constructed similarly to the train set used in classification VQA models, thus the predefined set of answers could cover the correct answers to a great extent), the results of our "fusion-brain" model are rather promising, considering the versatility and simplicity of the approach, and prove the need for further research.

Tab. 3. Private scores for different training strategies

training setup	C2C CodeBLEU	HTR Acc	ZsOD F1	VQA Acc	Overall
Single-task	0.123	0.533	0.193	0.307	1.156
ZsOD + VQA	–	–	**0.196**	0.313	–
HTR + ZsOD + VQA	–	0.566	**0.196**	0.325	–
Fusion	**0.132**	**0.587**	0.191	**0.327**	**1.237**

Tab. 4. Scores of SOTA models on private test sets

	C2C **CodeBLEU** PLBART [52]	**HTR** **Acc** Easter2 [53]	**ZsOD** **F1** MDETR [48]	**VQA** **Acc** (classification)
score	0.309	0.761	0.359	0.955

4. Emissions reduction

Recently, reporting energy and carbon metrics of training deep learning models has become common practice to promote energy-efficient research [54, 55]. In [56], the Machine Learning Emissions Calculator (ML CO2) is proposed, which estimates carbon emissions based on GPU type, hours spent on training, cloud provider, and region. This approach is very useful as it does not require reproducing the training process [57]. According to ML CO2, we estimate (see Table 5) that training the model in the fusion setup generates almost one third less CO2eq (carbon-dioxide equivalent) than when training in a single-task regime, thus proving multi-task learning to be more energy-efficient and climate-friendly.

Tab. 5. Total parameters summarized for all 4 tasks

Training setup	**Training time (hours)**	**Training params**	**CO2 (kg)**
Single-task	48.5	3,283,978,882	59.20
Fusion	**35**	**988,272,474**	**42.72**

Conclusion

In this paper we have presented the AI Journey 2021 Challenge called FusionBrain [58] – to the best of our knowledge, the first competition that is dedicated to the creation of a unified architecture which could deal with different modalities and solve 4 tasks for vision, language and programming code: Code2code Translation, Handwritten Text recognition, Zero-shot Object Detection, and Visual Question Answering. To test the participants' submissions, the datasets for each task were created; we also have described how the data were prepared. To date, the Russian part of the proposed dataset for HTR task is the largest Russian handwritten dataset in the world.

We also came up with a task statement and a competition design for the FusionBrain Challenge. Actually, there were 41 teams that took part in the competition and made at least one submission, and 513 submissions in total (refer to [59]).

We suppose that one of the main questions for future research in multimodal and multitask learning is how to combine tasks during the training so that the knowledge obtained by the model within one task would contribute to solving other tasks. It seems that some sort of hierarchical clustering of tasks is needed. According to our experiments, although C2C task is more different from all other tasks, adding it to the training procedure doesn't deteriorate the performance on other tasks (except for a slight drop in ZsOD score), but even improves the performance for HTR.

The results of the competition prove that using foundation models for solving several tasks in different modalities is promising as in all three prize solutions different foundation models (BART [36], T5 [37, 60], GPT) are used (see Appendix A). We believe that the results would be more strong if not for the lack of time given to the participants to solve such a challenging task (1 month). The task of creating a "universal" model, a multimodal and multitask architecture, is very deep and thus, we think that it's worth creating a benchmark and organizing a challenge on a long-term basis.

We believe that the FusionBrain Challenge became a successful first attempt to organize such a competition. The appearance of numerous new multimodal and multitask architectures (GATO, OFA, Flamingo, etc.) in the year following the competition proves the relevance and prospects of this topic. Experience gained during the holding of the competition allowed us to come up with a logical continuation – FusionBrain Challenge 2.0 in which we decided to focus on solving problems within two modalities that are naturally combined (text & image) – and the possibility of exploring unimodal and cross-modal connections. The novelties also relate to the formulation of tasks: in the case of the FusionBrain Challenge 2.0, it is made in a more natural way – with the tasks expressed in natural language and the inclusion of hidden tasks which are unknown to the participants.

Acknowledgements

We would like to thank Sber and SberCloud for granting the GPU-resources to us to experiment with different architectures and also to the participants to train their models, and for supporting the FusionBrain Challenge in general.

References

[1] Sludnova AA, Shutko VV, Gaidel AV, Pavel Mikhailovich Zelter PM, Kapishnikov AV, Nikonorov AV. Identification of pathological changes in the lungs using an analysis of radiological reports and tomographic images. Computer Optics 2021; 45(2): 261-266. DOI: 10.18287/2412-6179-CO-793.

[2] Liu X, He P, W, Gao J. Multi-task deep neural networks for natural language understanding. arXiv preprint. 2019. Source: ⟨https://arxiv.org/abs/1901.11504⟩.

[3] Hu R, Singh A. Unit: Multimodal multitask learning with a unified transformer. arXiv preprint. 2021. Siurce: ⟨https://arxiv.org/abs/2102.10772⟩.

[4] Liang PP, Liu Z, Zadeh AB, Morency L-P. Multimodal language analysis with recurrent multistage fusion. Proc 2018 Conf on Empirical Methods in Natural Language Processing 2018: 150-161.

[5] Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W. Visualbert: A simple and performant baseline for vision and language. arXiv preprint. 2019. Source: ⟨https://arxiv.org/abs/1908.03557⟩.

[6] Das A, Wahi JS, Li S. Detecting hate speech in multi-modal memes. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/2012.14891⟩.

[7] Savchenko A, Alekseev A, Kwon S, Tutubalina E, Myasnikov E, Nikolenko S. Ad lingua: Text classification

improves symbolism prediction in image advertisements. Proc 28th Int Conf on Computational Linguistics 2020: 1886-1892. DOI: 10.18653/v1/2020.coling-main.171.

[8] Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J. Perceiver: General perception with iterative attention. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2103.03206⟩.

[9] Jaegle A, Borgeaud S, Alayrac J-B, et al. Perceiver IO: A general architecture for structured inputs & outputs. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2107.14795⟩.

[10] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. Preprint. 2019. Source: ⟨https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf⟩.

[11] Alayrac J-B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T, Gong Z, Samangooei S, Monteiro M, Menick J, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Binkowski M, Barreira R, Vinyals O, Zisserman A, Simonyan K. Flamingo: a visual language model for few-shot learning. arXiv preprint. 2022. Source: ⟨https://arxiv.org/abs/2204.14198⟩.

[12] Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, de Las Casas D, Hendricks LA, Welbl J, Clark A, Hennigan T, Noland E, Millican K, van den Driessche G, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Rae JW, Vinyals O, Sifre L. Training compute-optimal large language models. arXiv preprint. 2022. Source: ⟨https://arxiv.org/abs/2203.15556⟩.

[13] Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C. Attention bottlenecks for multimodal fusion. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2107.00135⟩.

[14] Komkov S, Dzabraev M, Petiushko A. Mutual modality learning for video action classification. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/2011.02543⟩.

[15] Lu K, Grover A, Abbeel P, Mordatch I. Pretrained transformers as universal computation engines. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2103.05247⟩.

[16] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S. Parameter-efficient transfer learning for NLP. Proc 36th Int Conf on Machine Learning (PMLR '19) 2019: 2790-2799.

[17] Pfeiffer J, Kamath A, Rücklé A, Cho K, Gurevych I. AdapterFusion: Non-destructive task composition for transfer learning. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/2005.00247⟩.

[18] Tay Y, Zhao Z, Bahri D, Metzler D, Juan D-C. HyperGrid transformers: Towards a single model for multiple tasks. International Conference on Learning Representations (ICLR 2021) 2021: 1-14. Source: ⟨https://openreview.net/pdf?id=hiq1rHO8pNT⟩.

[19] Pilault J, Elhattami A, Pal C. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/2009.09139⟩.

[20] Reed S, Zolna K, Parisotto E, Colmenarejo SG, Novikov A, Barth-Maron G, Gimenez M, Sulsky Y, Kay J, Springenberg JT, Eccles T, Bruce J, Razavi A, Edwards A, Heess N, Chen Y, Hadsell R, Vinyals O, Bordbar M, de Freitas N. A generalist agent. arXiv preprint. 2022. Source: ⟨https://arxiv.org/abs/2205.06175⟩.

[21] Maillard J, Karpukhin V, Petroni F, Yih W-t, Ouguz B, Stoyanov V, Ghosh G. Multi-task retrieval for knowledge-intensive tasks. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2101.00117⟩.

[22] Gabeur V, Sun C, Alahari K, Schmid C. Multi-modal transformer for video retrieval. 16th European Conf on Computer Vision (ECCV 2020) 2020: 214-229.

[23] Dzabraev M, Kalashnikov M, Komkov S, Petiushko A. MDMMT: Multidomain multimodal transformer for video retrieval. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition 2021: 3354-3363.

[24] Ahmad WU, Tushar MdGR, Chakraborty S, Chang K-W. AVATAR: A parallel corpus for java-python program translation. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2108.11590⟩.

[25] Python Tokenizer. 2021. Source: ⟨https://docs.python.org/3/library/tokenize.html⟩.

[26] Puri R, Kung DS, Janssen G, Zhang W, Domeniconi G, Zolotov V, Dolby J, Chen J, Choudhury M, Decker L, Thost V, Buratti L, Pujar S, Ramji S, Finkler U, Malaika S, Reiss F. CodeNet: A large-scale ai for code dataset for learning a diversity of coding tasks. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2105.12655⟩.

[27] Ren S, Guo D, Lu S, Zhou L, Liu S, Tang D, Sundaresan N, Zhou M, Blanco A, Ma S. CodeBLEU: a method for automatic evaluation of code synthesis. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/2009.10297⟩.

[28] IAM handwriting database. 2021. Source: ⟨https://fki.tic.heia-fr.ch/databases/iam-handwriting-database⟩.

[29] HTRdataset. 2021. Source: ⟨https://github.com/sberbank-ai/htrdatasets⟩.

[30] Gu X, Lin T-Y, Kuo W, Cui Y. Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2104.13921⟩.

[31] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA, Bernstein MS, Li F-F. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint. 2016. Source: ⟨https://arxiv.org/abs/1602.07332⟩.

[32] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. arXiv preprint. 2015. Source: ⟨https://arxiv.org/abs/1503.01817⟩.

[33] FusionBrain Concept. 2021. Source: ⟨https://colab.research.google.com/drive/1YAkxWG0dRKPtqy9CZxFPvCNCCXvMGr65?usp=sharing⟩.

[34] Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2108.07258⟩.

[35] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2019. Source: ⟨https://arxiv.org/abs/1810.04805⟩.

[36] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint. 2019. Source: ⟨https://arxiv.org/abs/1910.13461⟩.

[37] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/1910.10683⟩.

[38] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T,

Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/2005.14165⟩.

[39] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2103.00020⟩.

[40] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2102.12092⟩.

[41] Rubinstein R, Davidson W. The cross-entropy method for combinatorial and continuous optimization. Methodol Comput Appl Probab 1999; 1: 127-190.

[42] Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets. Proc 23rd Int Conf on Machine learning (ICML'06) 2006: 369-376.

[43] Shonenkov A, Karachev D, Novopoltsev M, Potanin M, Dimitrov D. StackMix and blot augmentations for handwritten text recognition. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2108.11667⟩.

[44] de Buy Wenniger GM, Schomaker L, Way A. No padding please: Efficient neural handwriting recognition. 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 355–362. IEEE, 2019.

[45] Johannes Michael, Roger Labahn, Tobias Gru¨ning, and Jochen Zollner. Evaluating sequence-to-sequence models for handwritten text recognition. 2019 Int Conf on Document Analysis and Recognition (ICDAR) 2019: 1286-1293.

[46] Potanin M, Dimitrov D, Shonenkov A, Bataev V, Karachev D, Novopoltsev M. Digital Peter: Dataset, competition and handwriting recognition methods. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2103.09354⟩.

[47] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv preprint. 2015. Source: ⟨https://arxiv.org/abs/1512.03385⟩.

[48] Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I, Carion N. MDETR – modulated detection for end-to-end multi-modal understanding. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2104.12763⟩.

[49] Rezatofighi H, Tsoi N, Gwak JY, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2019: 658-666.

[50] Marti UV, Bunke H. The IAM-database: An english sentence database for offline handwriting recognition. Int J Doc Anal Recognit 2002; 5: 39-46.

[51] Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. arXiv preprint. 2016. Source: ⟨https://arxiv.org/abs/1612.00837⟩.

[52] Ahmad WU, Chakraborty S, Ray B, Chang K-W. Unified pre-training for program understanding and generation. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2103.06333⟩.

[53] Chaudhary K, Bali R. Easter2.0: Improving convolutional models for handwritten text recognition. arXiv preprint. 2022. Source: ⟨https://arxiv.org/abs/2205.14879?context=cs.AI⟩.

[54] Henderson P, Hu J, Romoff J, Emma B, Jurafsky D, Pineau J. Towards the systematic reporting of the energy and carbon footprints of machine learning. J Mach Learn Res 2020; 21(248): 1-43.

[55] Patterson D, Gonzalez J, Le Q, Liang C, Munguia L-M, Rothchild D, So D, Texier M, Dean J. Carbon emissions and large neural network training. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2104.10350⟩.

[56] Lacoste A, Luccioni A, Schmidt V, Dandres T. Quantifying the carbon emissions of machine learning. arXiv preprint. 2019. Source: ⟨https://arxiv.org/abs/1910.09700⟩.

[57] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit – Leveraging artificial intelligence to combat climate change: Opportunities, challenges, and recommendations. AI Soc 2021; 18: 1-25.

[58] FusionBrain challenge. 2021. Source: ⟨https://github.com/ai-forever/fusion_brain_aij2021⟩.

[59] DS Works. 2021. Source: ⟨https://dsworks.ru/champ/fb5778a8-94e9-46de-8bad-aa2c83a755fb⟩.

[60] Cho J, Lei J, Tan H, Bansal M. Unifying vision-and-language tasks via text generation. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2102.02779⟩.

[61] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers amp; distillation through attention. arXiv preprint. 2020. Source: ⟨https://arxiv.org/abs/2012.12877⟩.

[62] Li M, Lv T, Chen J, Cui L, Lu Y, Florencio D, Zhang C, Li Z, Wei F. TrOCR: Transformer-based optical character recognition with pre-trained models. arXiv preprint. 2021. Source: ⟨https://arxiv.org/abs/2109.10282⟩.

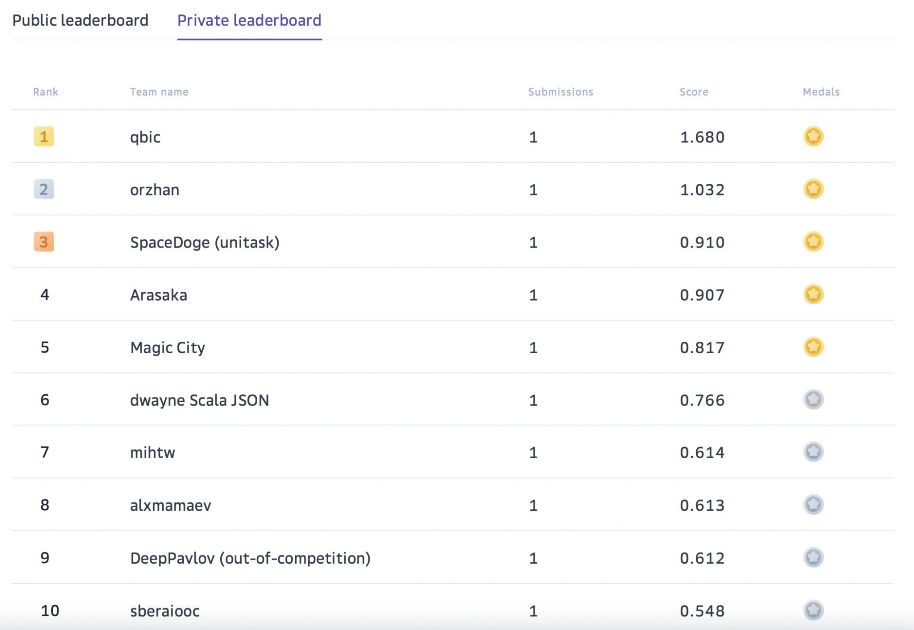
Appendix A. Private leaderboard

We provide the private leaderboard on the FusionBrain Challenge (see Figure 3).

Metrics of the winner of the competition can be seen in Table 6. The winner of the competition (*qbic*) uses BART [36] as the core of the architecture: for C2C task the encoder-decoder architecture is used as is; for HTR task, the input is first passed through pretrained DeiT [61] (as in TrOCR [62]) before moving to the BART's encoder, thus DeiT serves as an "adapter" that converts input visual data for this task to the BART encoder's familiar format; for crossmodal tasks (ZsOD, VQA), BART model is inserted into MDETR pipeline. The shared parameters of the whole architecture are 55.9 % of the number of all parameters.

The participant who took the second prize (*orzhan*) utilizes a pretrained VL-T5 model [60]: for vision-and-language tasks it is used as is (ZsOD task is set as image-text matching + visual grounding), for HTR and C2C tasks adapter layers are added.

Finally, the third-placed participant's (*Arasaka*) solution is based upon the baseline model (thus, GPT is used as the core) with several modifications in attention layers and loss functions.



Public leaderboard Private leaderboard

Rank	Team name	Submissions	Score	Medals
1	qbic	1	1.680	●
2	orzhan	1	1.032	●
3	SpaceDoge (unitask)	1	0.910	●
4	Arasaka	1	0.907	●
5	Magic City	1	0.817	●
6	dwayne Scala JSON	1	0.766	○
7	mihtw	1	0.614	○
8	alxmamaev	1	0.613	○
9	DeepPavlov (out-of-competition)	1	0.612	○
10	sberaiooc	1	0.548	○

Fig. 3. Top-10 participants sorted by Total score in private leaderboard

Tab. 6. Top-3 private scores of the multi-modal and multi-task models provided by participants of the FusionBrain Challenge

name	**CodeBLEU**	**Acc**	**F1**	**Acc**	**Total**
qbic	0.320	0.744	0.250	0.365	1.680
orzhan	0.233	0.314	0.166	0.318	1.032
Arasaka	0.218	0.377	0.074	0.237	0.907

Authors' information

Daria Dmitrievna Bakshandaeva is a researcher at the University of Helsinki. She holds a master's degree in Computational Linguistics from HSE University, Moscow. She is the author of papers accepted to the major NLP conferences (ACL, EMNLP). Her master's thesis considers different knowledge distillation approaches for two semantic similarity tasks: paraphrase identification and dialog response prediction. Currently, the main focus of her research is multimodal and multitask neural network architectures. E-mail: *daria.bakshandaeva@helsinki.fi* .

Denis Valerievich Dimitrov is a CV & Multimodal Research Lead at Sber AI and a Researcher of the Department of Probability Theory, Faculty of Mechanics and Mathematics at Lomonosov Moscow State University. His research interests include both strictly mathematical issues concerning statistical estimation of the f-divergences and applications such as multivariate inhomogeneities detection, feature selection, handwritten text recognition, generative computer vision models and multimodal models. He has a number of publications indexed in Scopus, including publications in Q1 journals such as Mathematics, Acta Mathematica Sinica and at such conferences as ICML, IJCAI, ICDAR etc.
E-mail: *Dimitrov.D.V@sberbank.ru .*

Vladimir Sergeyevich Arkhipkin is a fellow in SberAI and a second-year master's student at Moscow Institute of Physics and Technology (MIPT). His specializations are theoretical physics and machine learning, currently mainly interested in multimodal neural networks area of deep learning research. E-mail: *arkhipkin.v98@gmail.com .*

Alex Vladimirovich Shonenkov graduated from the Moscow Institute of Physics and Technology in 2018. Now he works as an ML researcher at SberAI; also, he is an AI enthusiast at Kaggle. Research interests: deep learning modality fusion, image processing, natural language processing, AI competitions. E-mail: *AVShonenkov@sberbank.ru* .

Mark Stanislavovich Potanin is a PhD student at the Moscow Institute of Physics and Technology, School of Applied Mathematics and Informatics. He works as a Machine Learning Engineer at MetaQuotes (Cyprus). His research interests include computer vision and neural architecture search. E-mail: *mark.potanin@phystech.edu* .

Denis Konstantinovich Karachev graduated from the Ural State University of Railway Transport, Faculty of Mechatronics and Robotics at 2016. He received a master's degree in Mechatronics and Robotics in 2018, and now he is a PhD student (Thesis: "Intelligent control system for an autonomous vehicle using neural network technologies"). He works in the position of a Senior researcher at OCRV, in the computer vision group. His research interest lies in different computer vision tasks. E-mail: *welcomedenk@gmail.com* .

Andrey Vladimirovich Kuznetsov, Ph.D., is a CV Lead at Sber AI and an Associate Professor at Samara National Research University. His research interests include image processing and machine learning algorithms development in remote sensing data analysis, digital image forgery detection, general CV tasks like classification, object detection and segmentation, etc. He has 56 publications indexed in Scopus, including publications in Q1 journals and on such conferences as ICPR, ICIAR, etc. He has a Top Rated Upwork profile with more than 20 successfully completed projects. He was a grantee and a lead researcher in several individual and group scientific projects funded by the Russian science foundations: RSF and RFBR. He developed a lecture and lab course on Technology of Secure Distributed Applications Development. He was a recipient of the Presidential Scholarship for Young Scientists in 2015-2017 and 2018-2020 and payments to young scientists and designers in Samara Region from 2015 to 2021.
E-mail: *AVladimirKuznetsov@sberbank.ru* .

Anton Dmitrievich Voronov is a research fellow in AIRI and a first-year PhD student at the Moscow Institute of Physics and Technology (MIPT). His major is Applied Mathematics and Physics, currently mainly interested in NLP and multimodal neural models areas of deep learning research. He has one publication accepted at arxiv org/abs/2109.08914EMNLP- 21. E-mail: *voronov@airi.net* .

Aleksandr Alexandrovich Petiushko graduated from Lomonosov Moscow State University, Faculty of Mechanics and Mathematics at 2006, and received an academic degree of a Candidate of Physical and Mathematical sciences at 2016 (Thesis: "Bigram languages discrete mathematics and mathematical cybernetics). He works in the role of a Leading researcher at Artificial Intelligence Research Institute (AIRI), leading the group of FusionBrain. His research interest lies in the application of empirical and theoretical robustness techniques to different tasks.
E-mail: *petiushko@airi.net* .

Vera Fedorovna Davydova is a researcher at Sber AI. She received a master’s degree in Computational linguistics from Higher School of Economics, Moscow. Her research interests include medical applications of natural language processing, word sense induction and text-to-code translation. Her works were accepted to top-NLP conferences (EMNLP, COLING). She also co-organized Social Media Mining Workshop at COLING 2022 conference.
E-mail: *veranchos@gmail.com* .

Elena Viktorovna Tutubalina, Ph.D. is a NLP Lead at Sber AI and AIRI. She is also an Associate Professor at Kazan Federal University. She is a grantee and a lead researcher in several group scientific projects funded by the Russian science foundations: RSF and RFBR. She has a number of publications indexed in Scopus, including publications in Q1 journals and world-leading NLP conferences (ACL, EMNLP, ECIR etc.). Her areas of interest include medical applications of natural language processing, information retrieval, and entity linking.
Email: *tutubalinaev@gmail.com* .

Received September 08, 2022. The final version – November 21, 2022.