AIRI

Trends in AI:
Multi-**modality**,
Multi-**tasking**,
Effectiveness

Aleksandr Petiushko

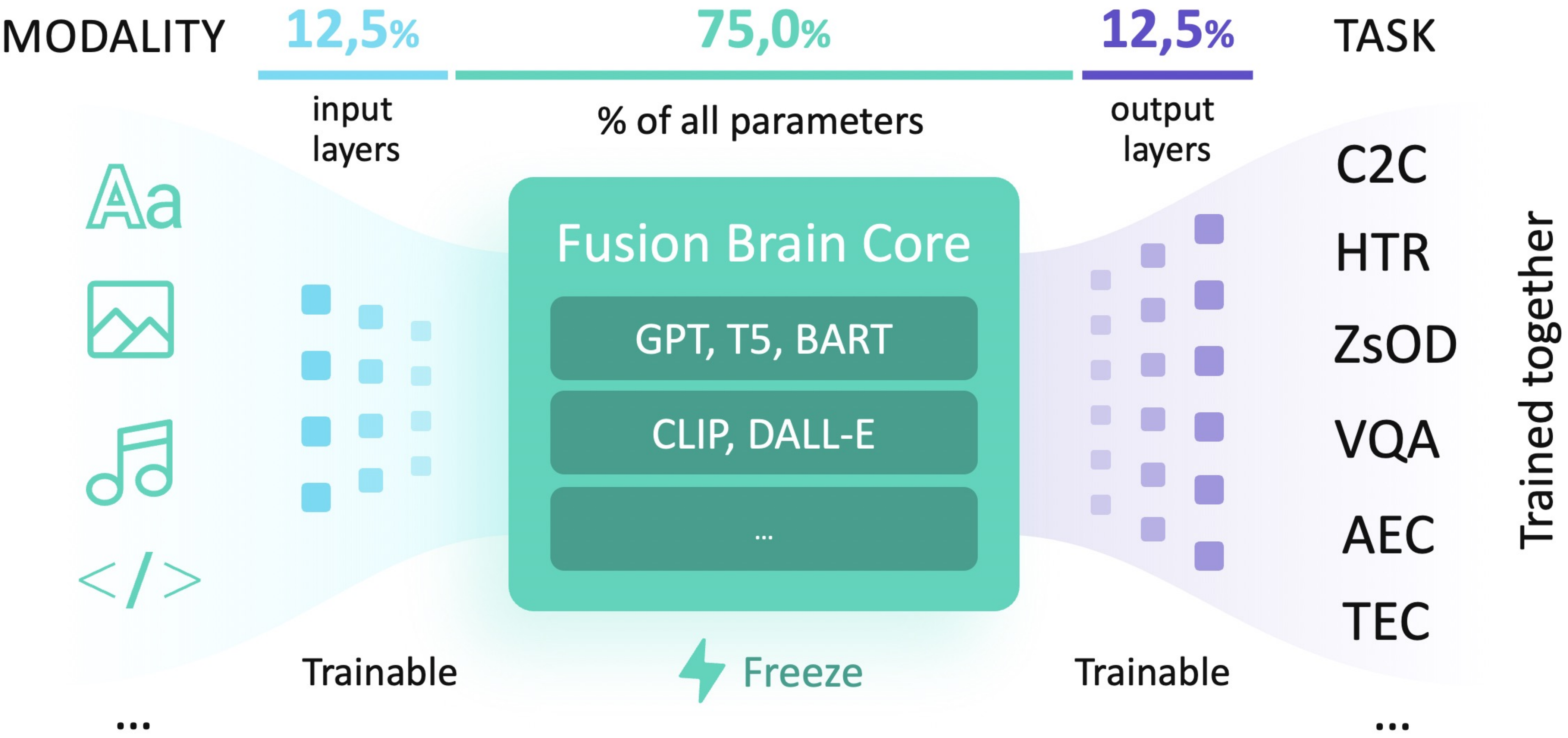
Leading scientific researcher, AIRI

January 27, 2022

01

Motivation

Motivation


MODALITY
12,5%
75,0%
12,5%
TASK
input layers
% of all parameters
output layers
Aa
Fusion Brain Core
GPT, T5, BART
CLIP, DALL-E
...
C2C
HTR
ZsOD
VQA
AEC
TEC
Trained together
Trainable
Freeze
Trainable
...
...

Motivation: business part

Problem
Retraining model from scratch is $$$

Separate training

- Generation of good text description
- Zero-shot object detection
- Handwritten text recognition
- Code2Code
- Visual Q&A
- …

Totally: ~ **$$$** M

Solution
Fine-tuning large pretrained model $$

Single pre-training

- GPT-3, DALL-E, CLIP

Separate fine-tuning

- Generation of good text description
- Zero-shot object detection
- Handwritten text recognition
- Code2Code
- Visual Q&A
- …

Totally: ~ **$$** M

Motivation: ecological part

Problem
Retraining model from scratch: CO_2 ↑↑↑

Solution
Fine-tuning large pretrained model: CO_2 ↑↑

Separate training
- Generation of good text description
- Zero-shot object detection
- Handwritten text recognition
- Code2Code
- Visual Q&A
- …

Totally: ~ **XXX kg CO_2e**

Single pre-training
- GPT-3, DALL-E, CLIP

Separate fine-tuning
- Generation of good text description
- Zero-shot object detection
- Handwritten text recognition
- Code2Code
- Visual Q&A
- …

Totally: ~ **XX kg CO_2e**

Motivation: trends

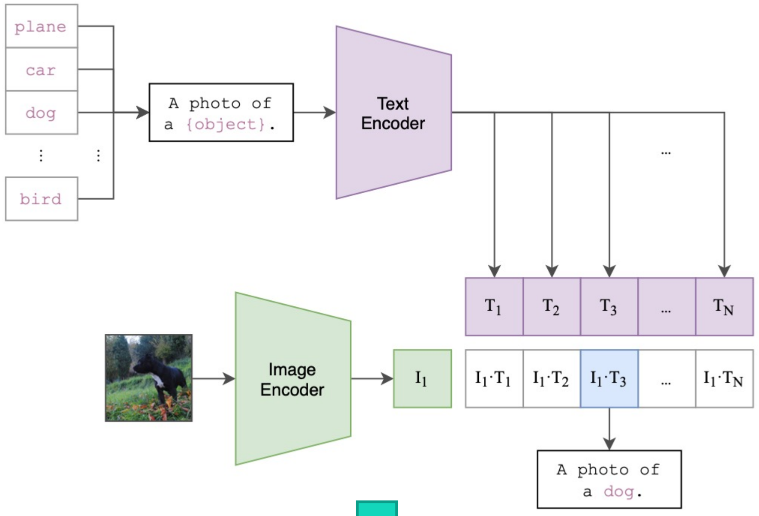
Current trends:

- Large pre-trained models (BERT, GPT-3)
- Multi-modality and multi-tasking (CLIP, DALL-E, UniT)

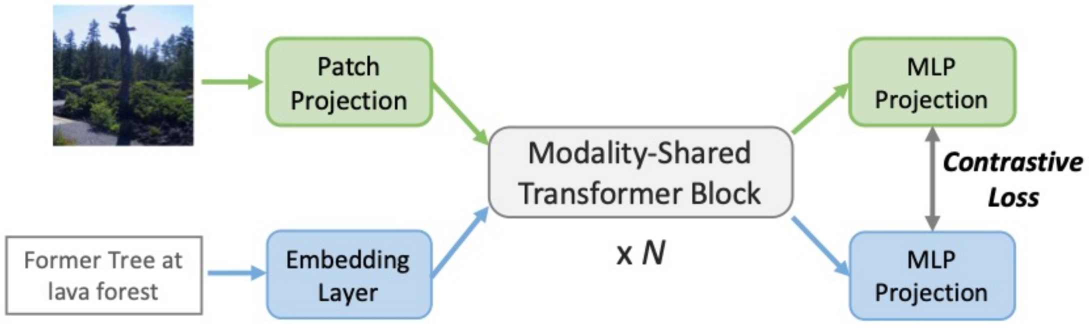
Target modalities: texts, images, sounds and other modalities like videos, programming languages, graphs and time series

Target tasks: NLP, CV and combined tasks like VQA

CLIP, 2021


plane
car
dog
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

MA-CLIP, 2022, ICLR

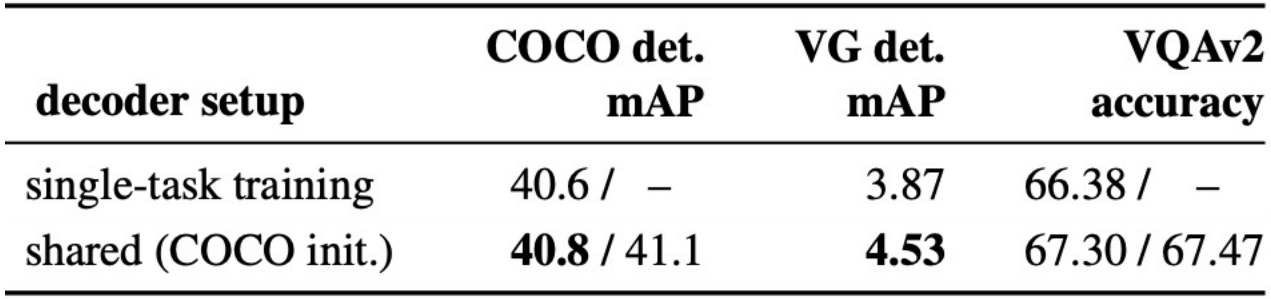


[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." 2021 (*OpenAI*).
[2] You, Haoxuan, et al. "MA-CLIP: Towards Modality-Agnostic Contrastive Language-Image Pre-training." 2021 (*Withdrawn submission to ICLR-2022*)

Motivation: WHY it is reasonable

Efficient multi-modality multi-task models

WHY we need multi-*



decoder setup	COCO det. mAP	VG det. mAP	VQAv2 accuracy
single-task training	40.6 / –	3.87	66.38 / –
shared (COCO init.)	**40.8** / 41.1	**4.53**	67.30 / 67.47

WHY we need efficiency

Model	#Params
GPT-3	175 B
Retrieval-based models	1 B (3*BERT-Large)

Motivation: WHY it is still non-solved

Model	#params	GLUE	SuperGLUE
RoBERTa-Large ST	8,5B	88.2	76.5
RoBERTa-Large MTL	355M	86.0	78.6
CA-MTL (RoBERTa-Large)	397,6M	**89.4**	**80.0**

Encoder (BERT)-based
Multi-task: **better**

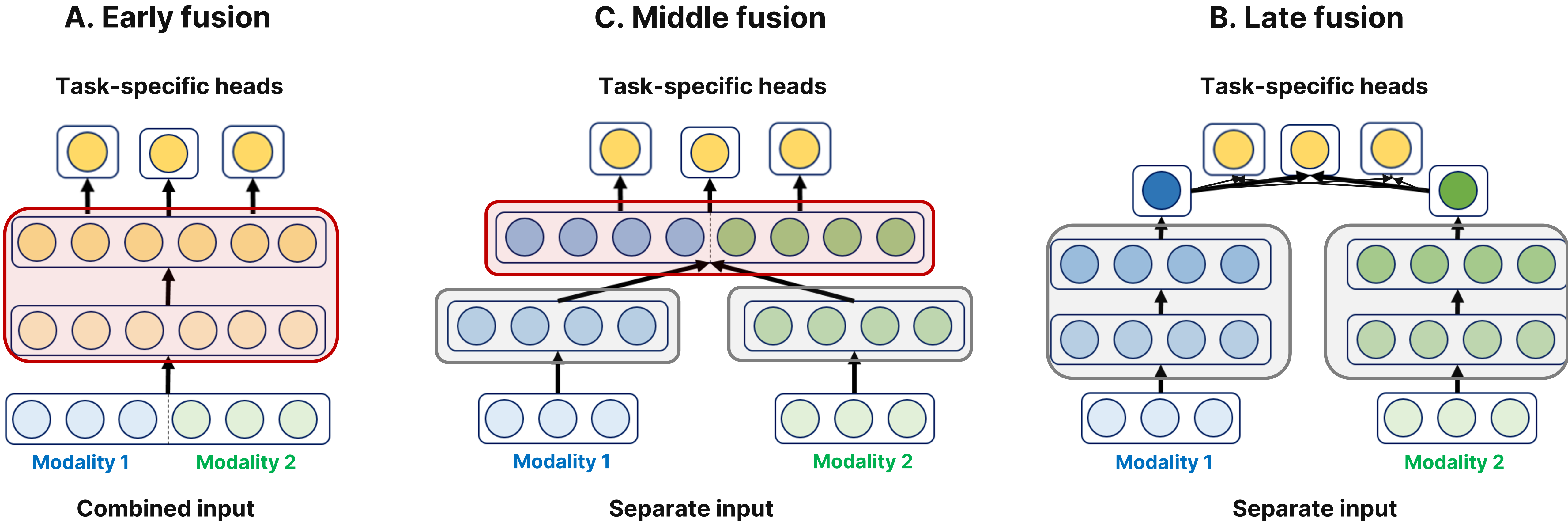
T5 (3B) STL	48B	88.5	86.4
HyperGrid (3B) MTL	3B	88.2	84.7
T5 (11B) STL	176B	**89.7**	**88.9**
HyperGrid (11B) MTL	11B	89.4	87.7

Decoder (T5)-based
Single-task: **better**

02

Multi-modality

Multi-modality: concepts



A. Early fusion

Task-specific heads

Combined input

C. Middle fusion

Task-specific heads

Separate input

B. Late fusion

Task-specific heads

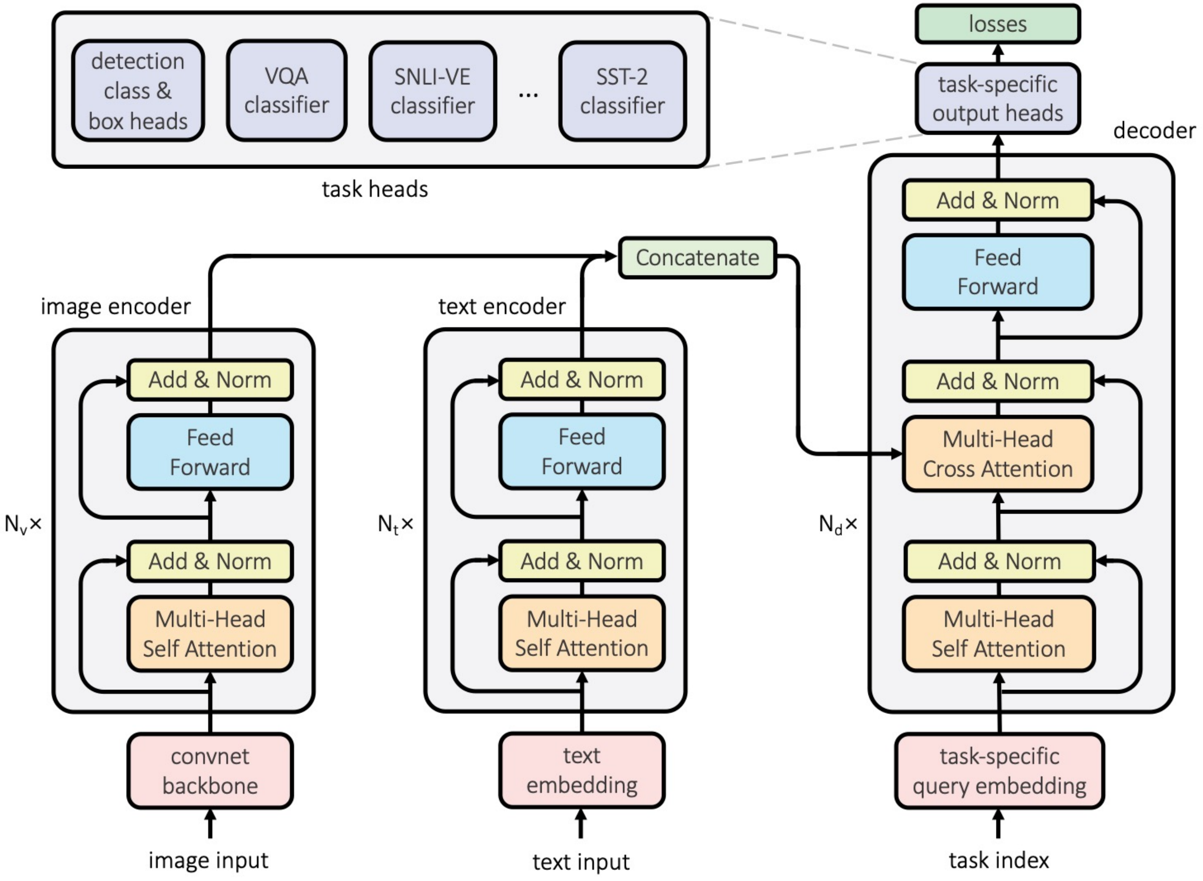
Separate input

🟥 Fusion processing

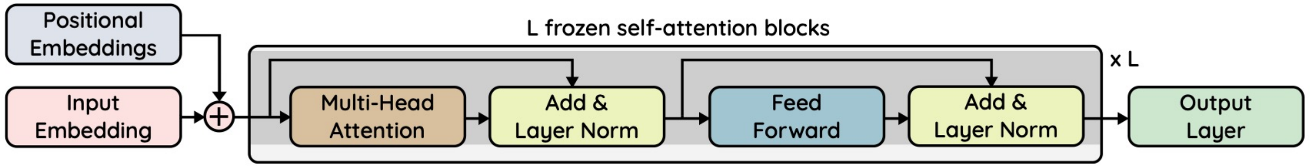
⬜ Modality-specific processing

Multi-modality: current trends

UniT[1]: via cross-attention



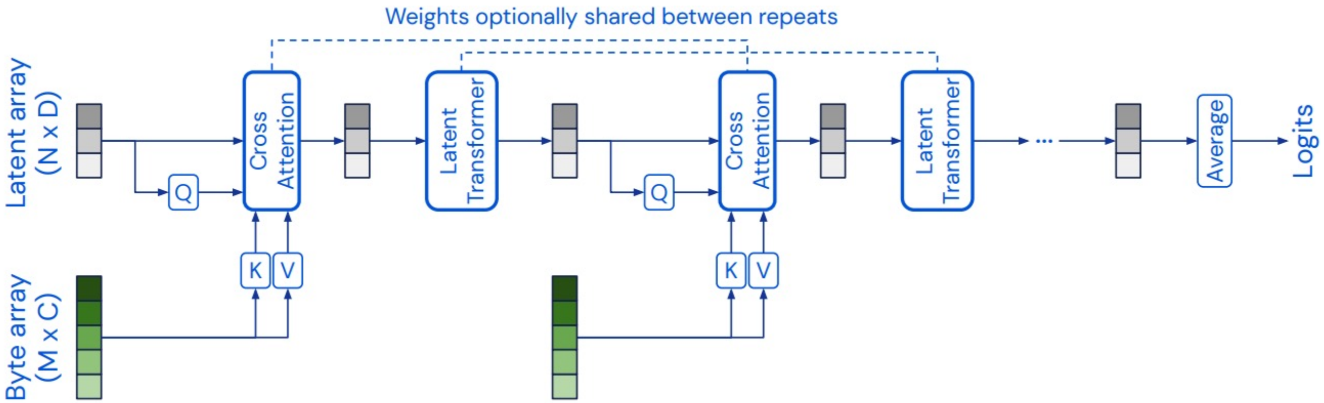
FPT[2]: via frozen MHA/FFN, tunable LN



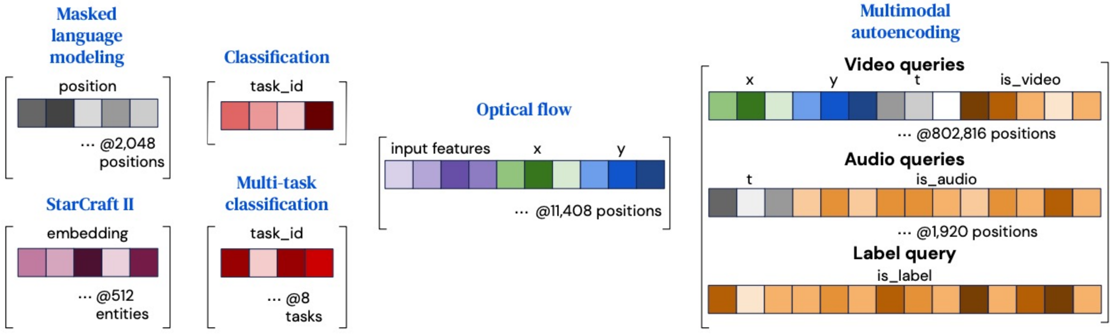
[1] Hu, Ronghang, and Amanpreet Singh. "UniT: Multimodal Multitask Learning with a Unified Transformer." 2021 (*Facebook*).
[2] Lu, Kevin, et al. "Pretrained transformers as universal computation engines." 2021 (*Google*)

Multi-modality: current trends

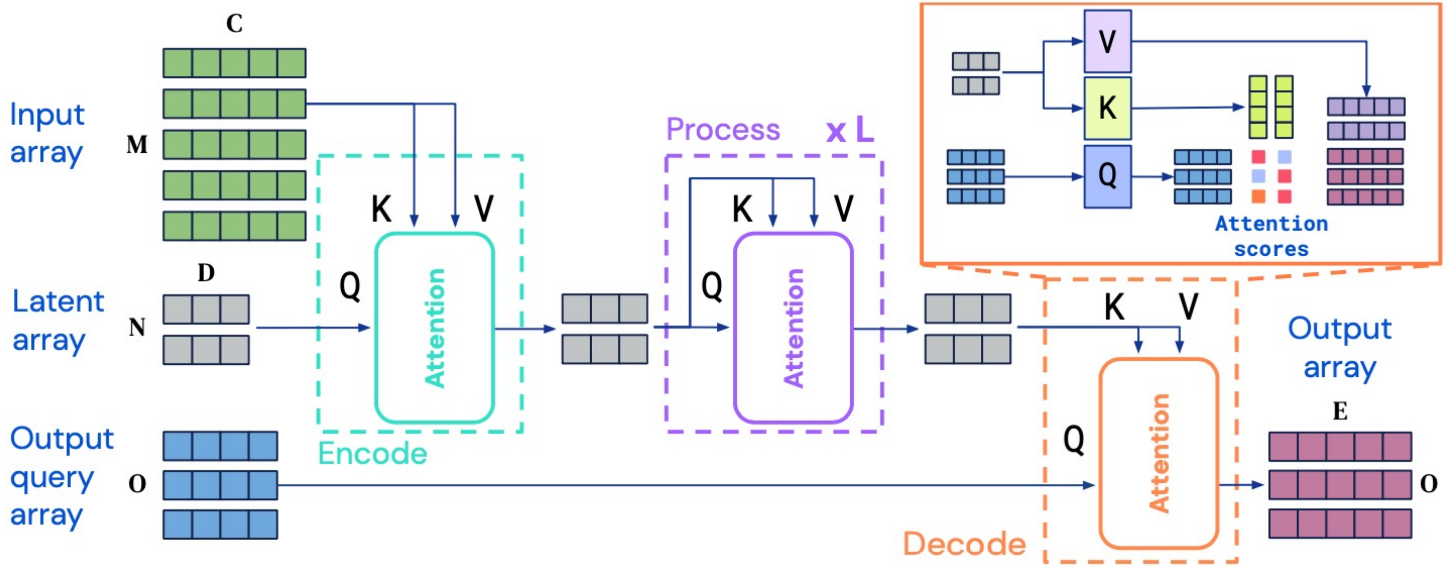
Perceiver[1]: iterative CA



Perceiver IO: output queries



Perceiver IO[2]**:** CA on input/output



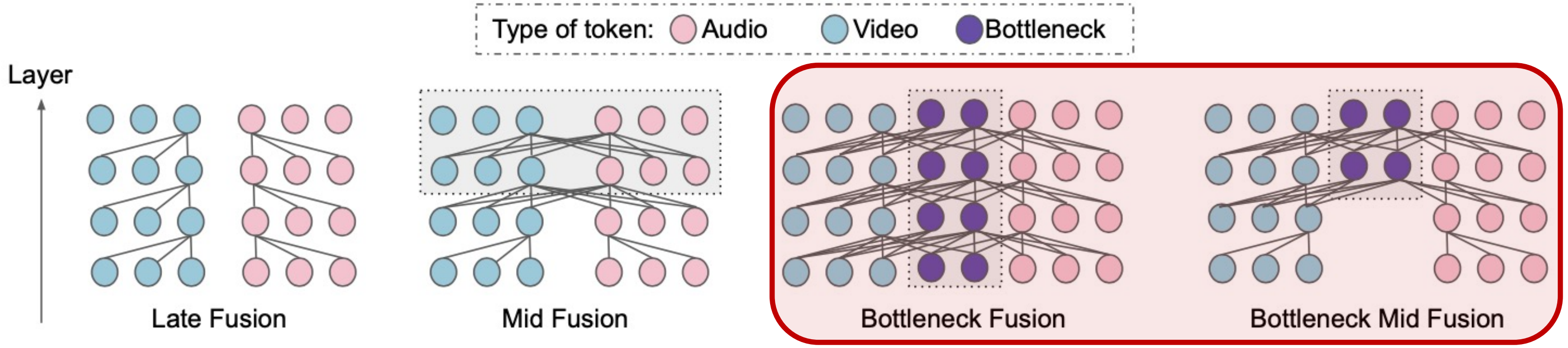
Main idea:

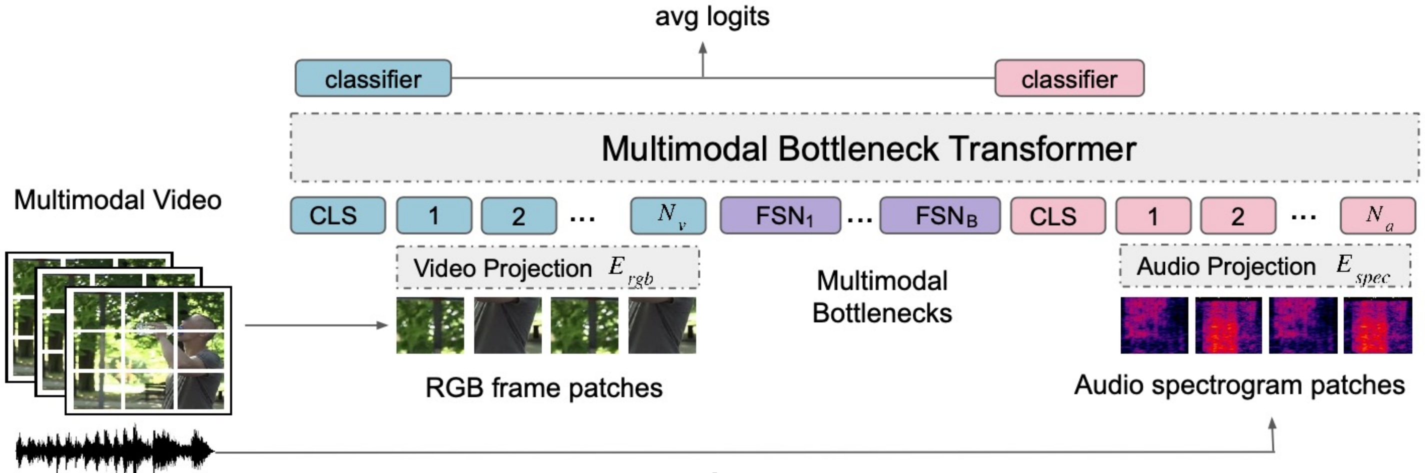
- **Iterative fusion** through **cross-attention** (query – latents, KV - input) allowing **linear** scaling on **input** size (not quadratic)
- Latent transformer is **GPT-2** like
- Weights of CA/SA are **shared**
- **Perceiver IO**[2] added ability to work with **multi-task and different output sizes** via CA where query is output structure, KV – latents (**complexity** – still the **linear** depending on the **output** size)

[1] Jaegle, Andrew, et al. "Perceiver: General perception with iterative attention." 2021 (*DeepMind*)

[2] Jaegle, Andrew, et al. "Perceiver io: A general architecture for structured inputs & outputs." 2021 (*DeepMind*)

Multi-modality: through information bottleneck



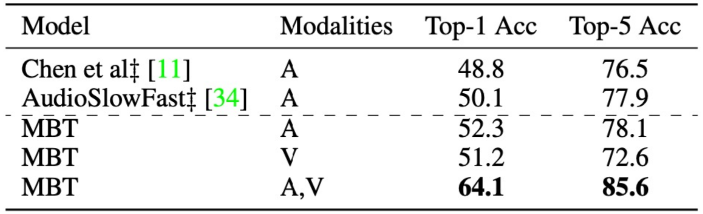


MBT[1] scheme

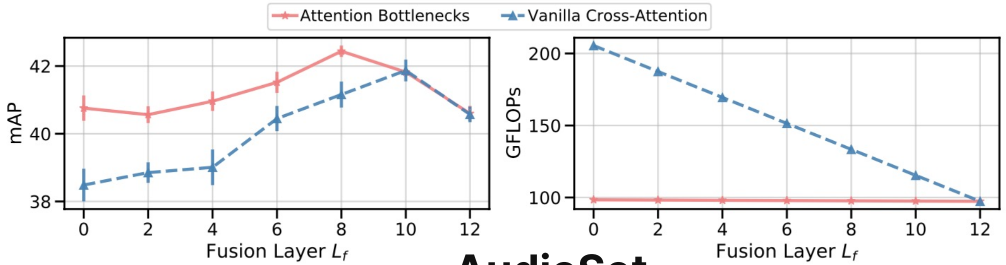
Main idea:

- **Middle-fusion through a small bottleneck** (B = 4 is used)
- **Fusion** is needed **closely to the top**

VGGSound[2]



Model	Modalities	Top-1 Acc	Top-5 Acc
Chen et al‡ [11]	A	48.8	76.5
AudioSlowFast‡ [34]	A	50.1	77.9
MBT	A	52.3	78.1
MBT	V	51.2	72.6
MBT	A,V	**64.1**	**85.6**



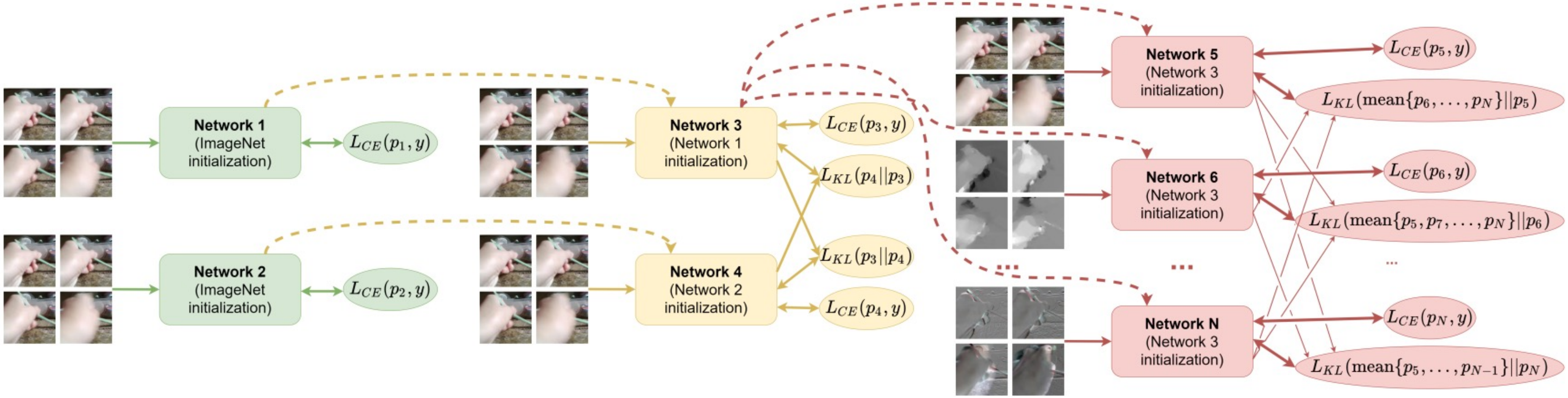
AudioSet

[1] Nagrani, Arsha, et al. "Attention Bottlenecks for Multimodal Fusion." 2021 (*Google*)
[2] https://www.robots.ox.ac.uk/~vgg/data/vggsound/

Multi-modality: through mutual learning

Main idea:

- **Pseudo multi-modality** through incorporation of knowledge by **mutual learning** technique
- **RGB** and **OpticalFlow** modalities for video action recognition were used


Network 1
(ImageNet initialization)
$L_{CE}(p_1, y)$
Network 2
(ImageNet initialization)
$L_{CE}(p_2, y)$
Network 3
(Network 1 initialization)
$L_{CE}(p_3, y)$
$L_{KL}(p_4||p_3)$
$L_{KL}(p_3||p_4)$
Network 4
(Network 2 initialization)
$L_{CE}(p_4, y)$
...
Network 5
(Network 3 initialization)
$L_{CE}(p_5, y)$
$L_{KL}(\text{mean}\{p_6, \dots, p_N\}||p_5)$
Network 6
(Network 3 initialization)
$L_{CE}(p_6, y)$
$L_{KL}(\text{mean}\{p_5, p_7, \dots, p_N\}||p_6)$
...
...
Network N
(Network 3 initialization)
$L_{CE}(p_N, y)$
$L_{KL}(\text{mean}\{p_5, \dots, p_{N-1}\}||p_N)$

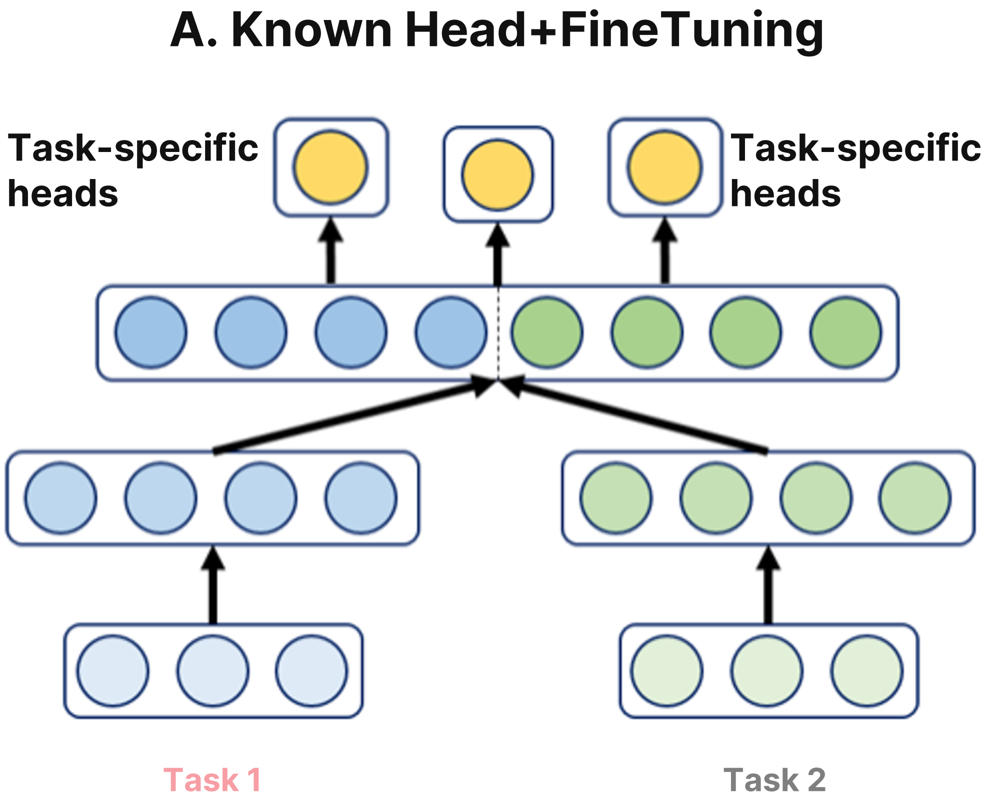
MML[1] scheme

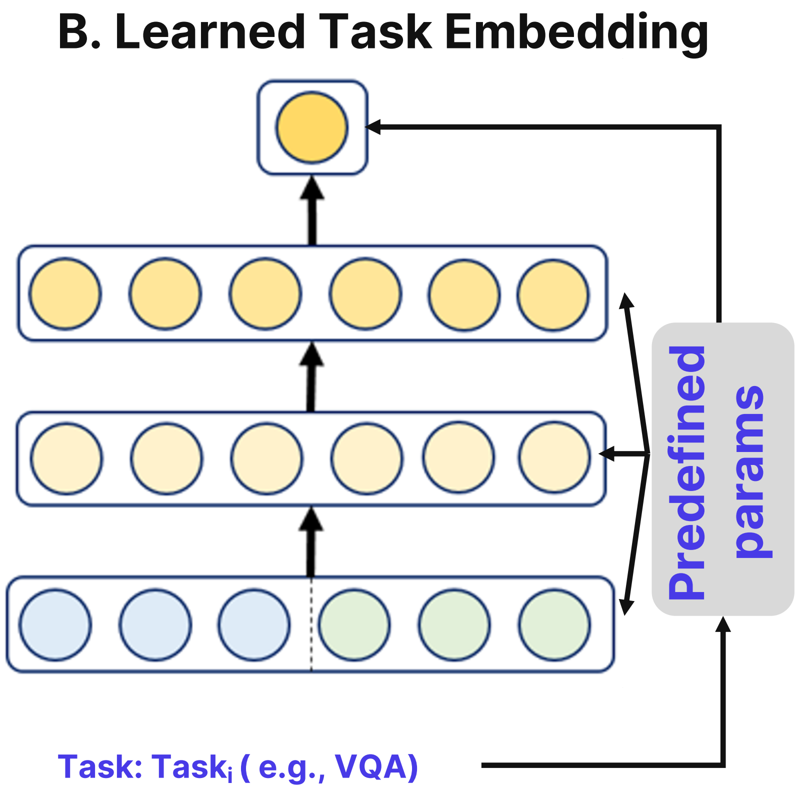
[1] Komkov, Stepan, Maksim Dzabraev, and Aleksandr Petiushko. "Mutual Modality Learning for Video Action Classification." 2020 (*Huawei*)

03

Multi-tasking

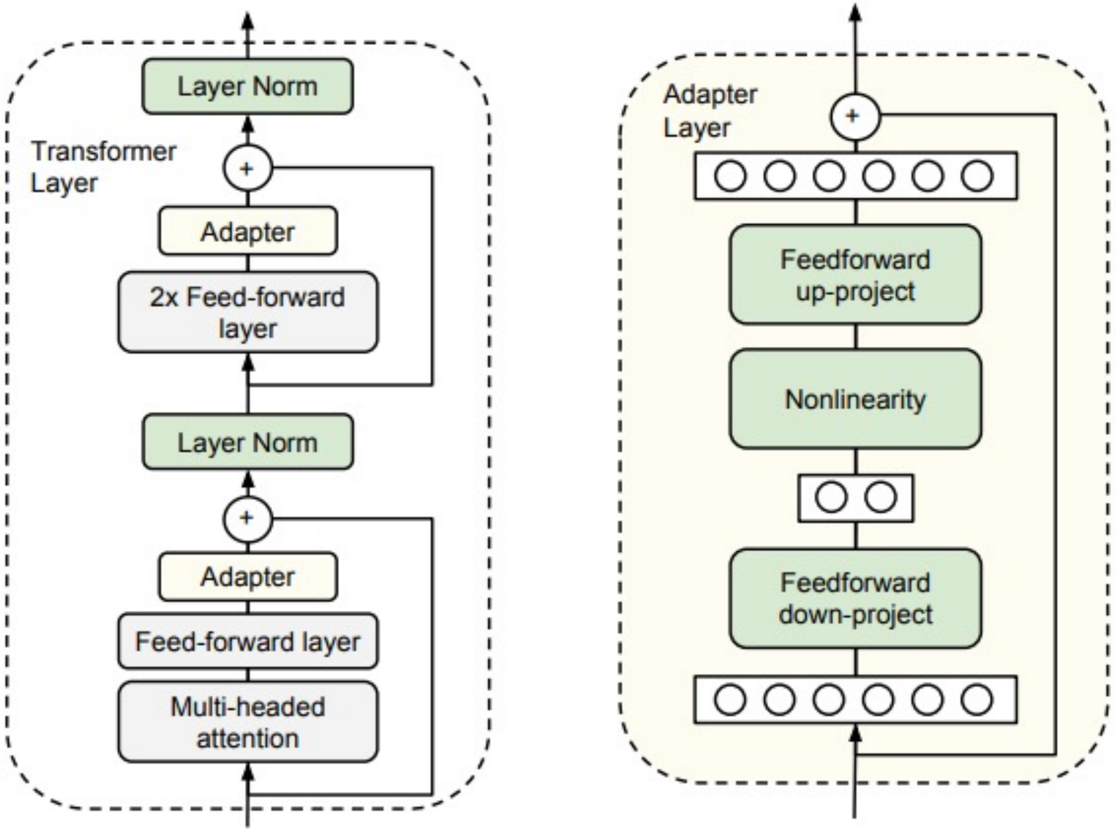
Multi-tasking: concepts





Multi-tasking: current trends

Adapters[1]: via task-specific learnable modules



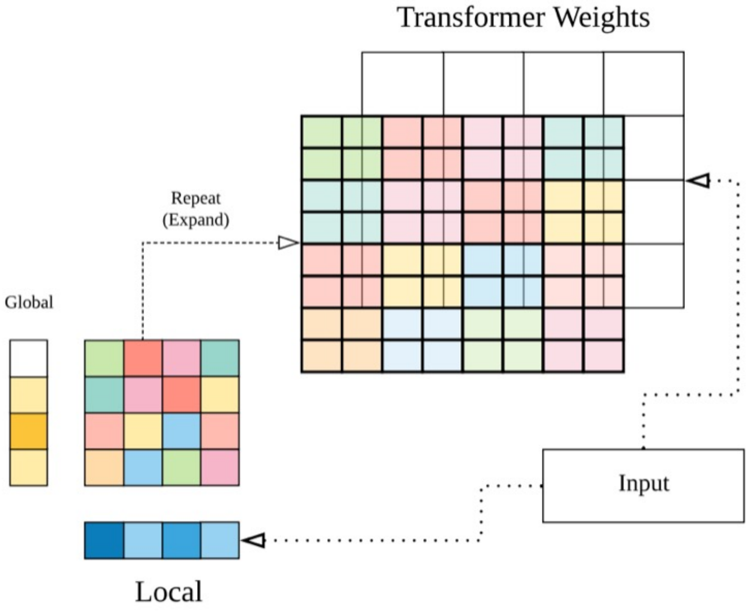
Main idea:

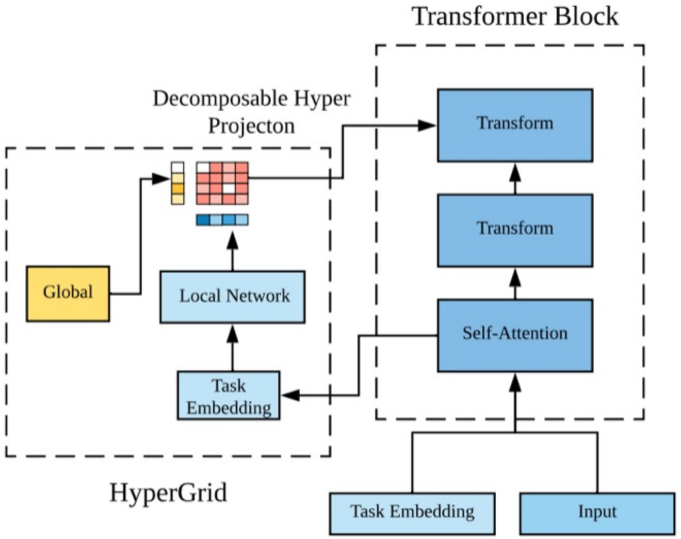
- **Freeze** the **Transformer** weights
- Add a small **learnable** task-specific module - **adapter**
- Performance close to single-task training, but **only +3.5% weights** for multi-task

[1] Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." 2019 (*Google*)

Multi-tasking: current trends

HyperGrid[1]: via dynamical weight matrix adjustment by learned task embedding

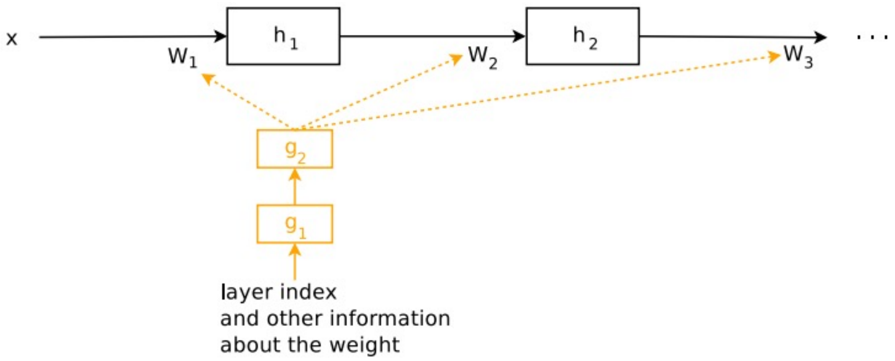




Main idea:

- **Learned task embedding** used to construct transformer matrix
- Back-bone transformer is T5
- Idea borrowed from **HyperNets**[2] conception

HyperNet concept

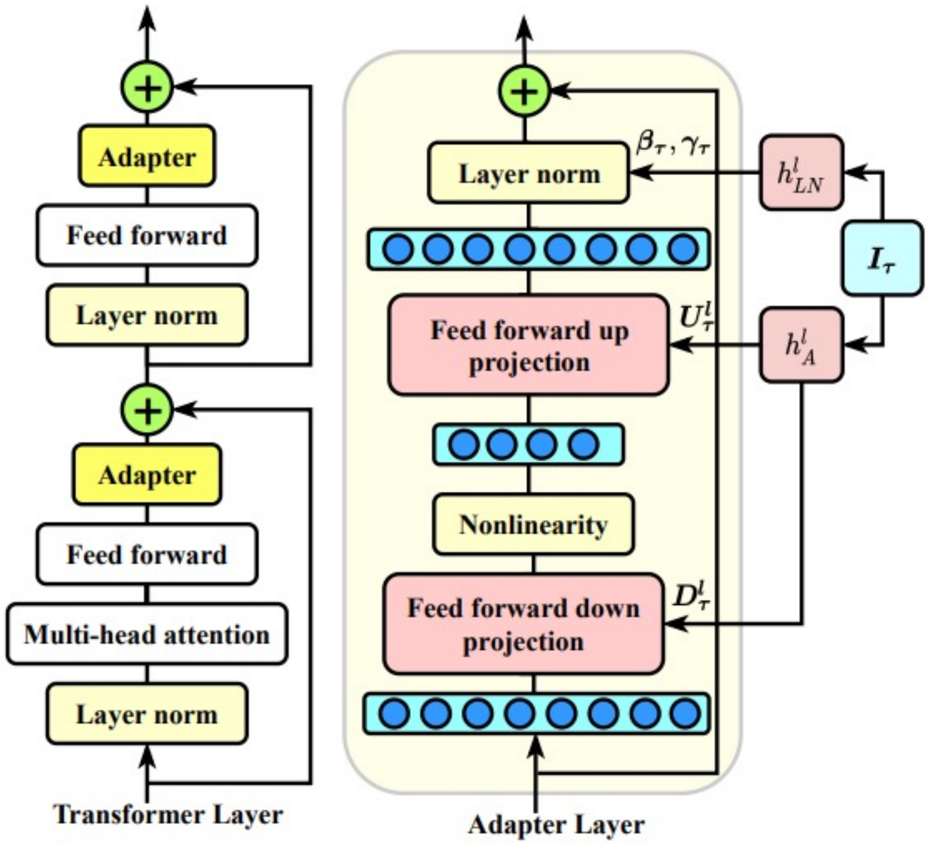


[1] Tay, Yi, et al. "HyperGrid Transformers: Towards A Single Model for Multiple Tasks." 2020 (*Google*)

[2] Ha, David, Andrew Dai, and Quoc V. Le. "Hypernetworks." 2016 (*Google*)

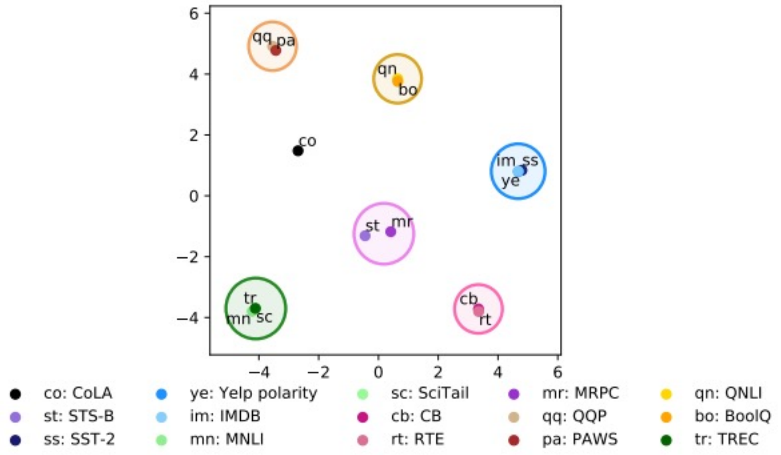
Multi-tasking: current trends

HyperFormer[1]: Adapters + HyperNets



Main idea:

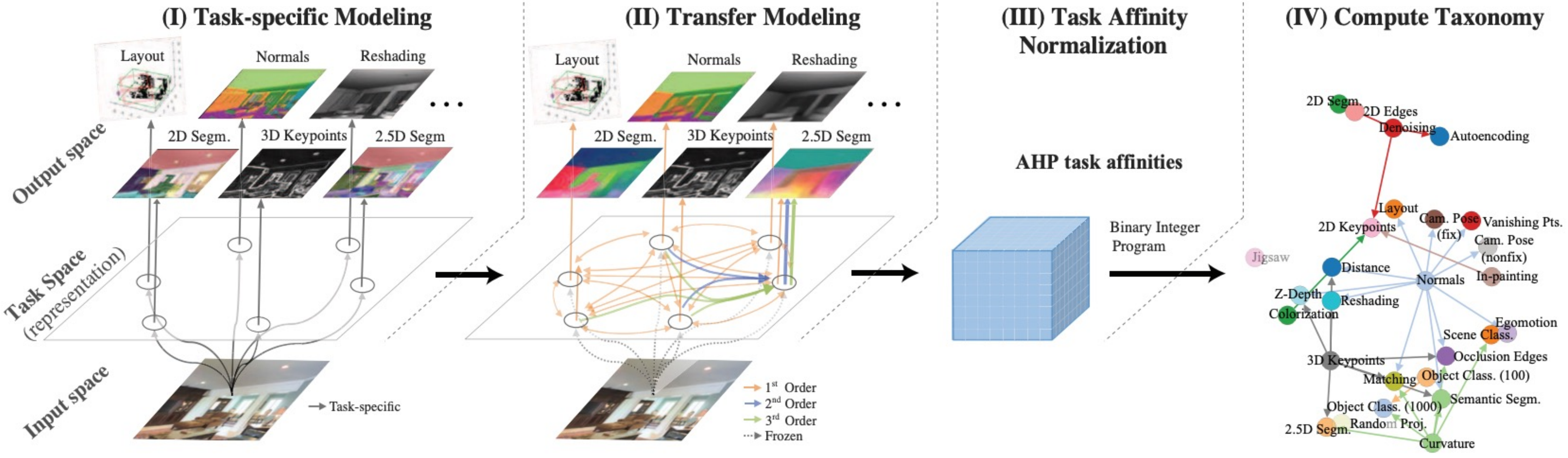
- Making the **Adapters** parameters *through* **HyperNets**
- New SotA with even **less** params than Adapters
- NLP task embeddings clusterization

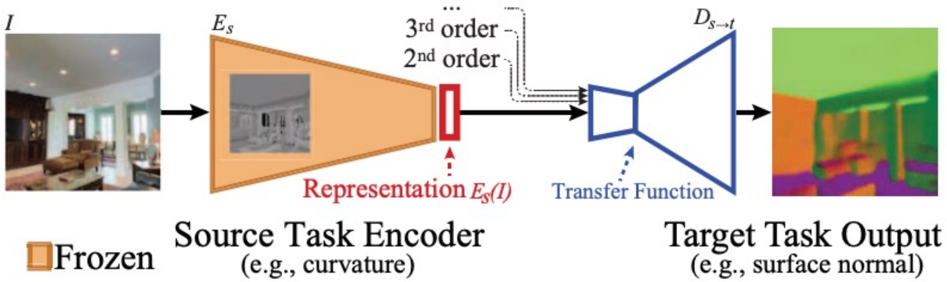


[1] Mahabadi, Rabeeh Karimi, et al. "Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks." 2021 (*Google*)

Multi-tasking: taskonomy

Taskonomy[1]: Task grouping via pairwise transfer performance

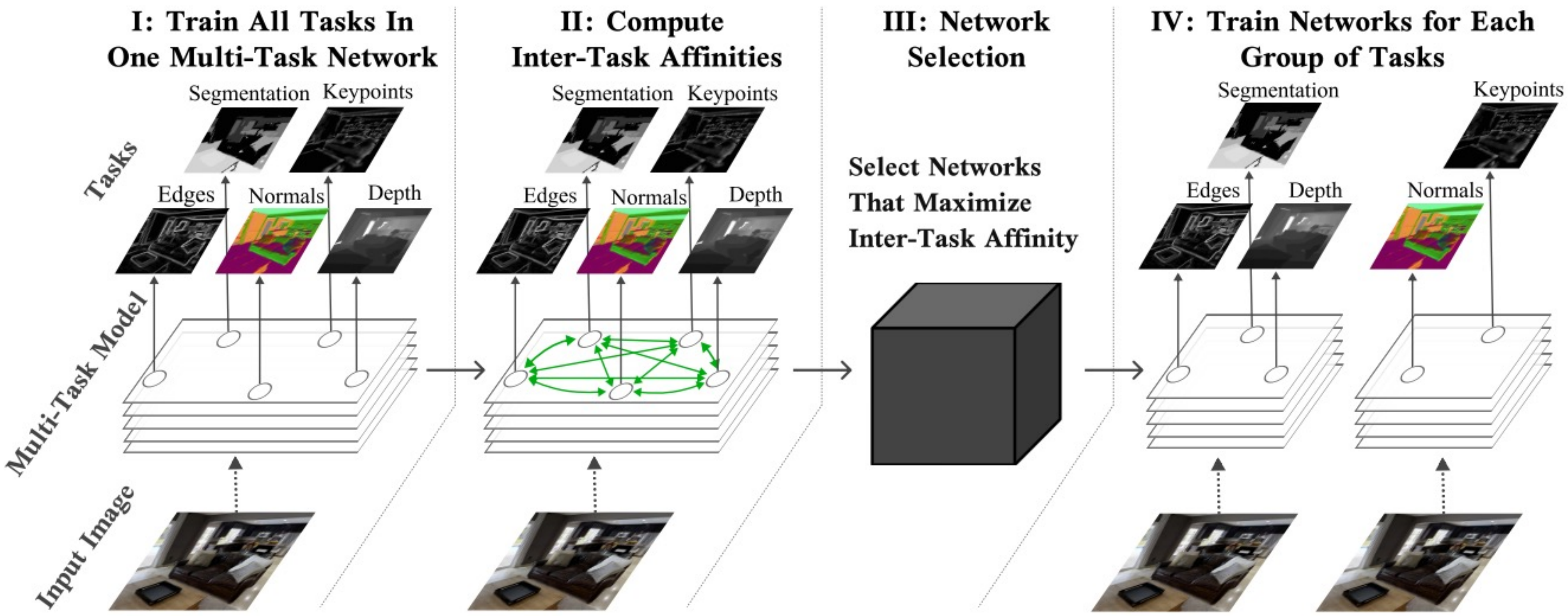




[1] Zamir, Amir R., et al. "Taskonomy: Disentangling task transfer learning." 2018 (*Stanford*)

Multi-tasking: how to group tasks

TAG[1]: Task grouping via similar gradient update

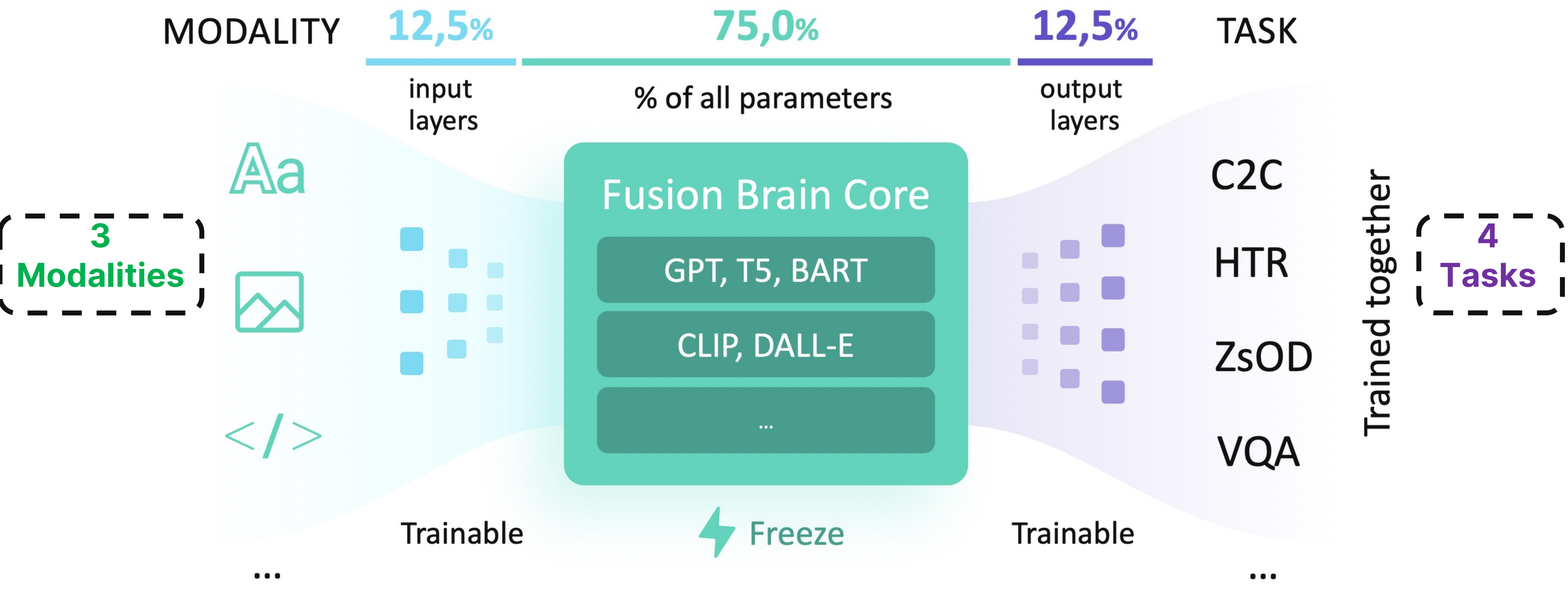

I: Train All Tasks In One Multi-Task Network
II: Compute Inter-Task Affinities
III: Network Selection
IV: Train Networks for Each Group of Tasks
Tasks
Multi-Task Model
Input Image
Segmentation
Keypoints
Edges
Normals
Depth
Segmentation
Keypoints
Edges
Normals
Depth
Select Networks That Maximize Inter-Task Affinity
Segmentation
Keypoints
Edges
Depth
Normals

[1] Fifty, Christopher, et al. "Efficiently identifying task groupings for multi-task learning." 2021 (*Google*)

04

Fusion Brain approach

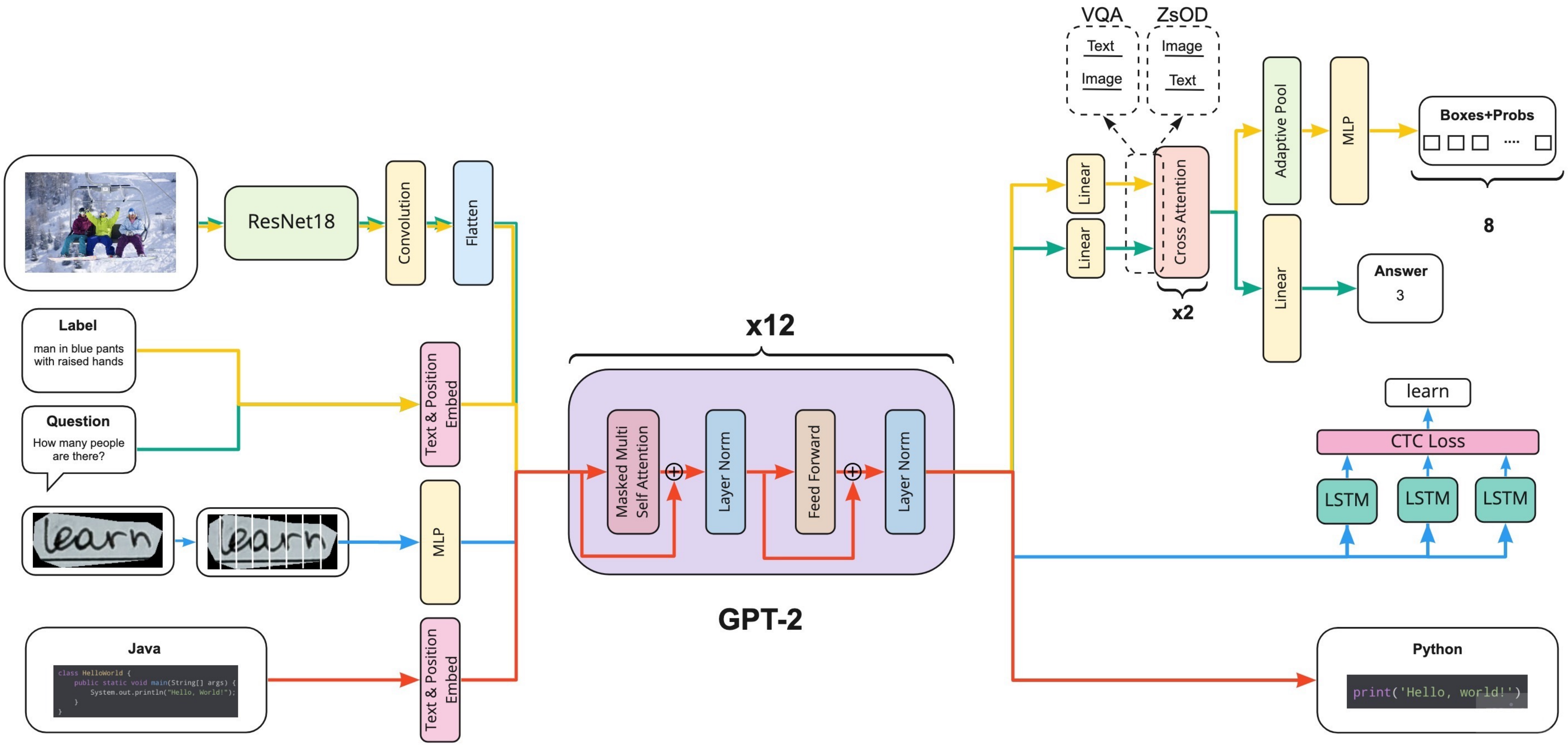
Fusion Brain concept[1]: overview



AIRI

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

Fusion Brain approach[1]: FPT, GPT-2, cross-attention



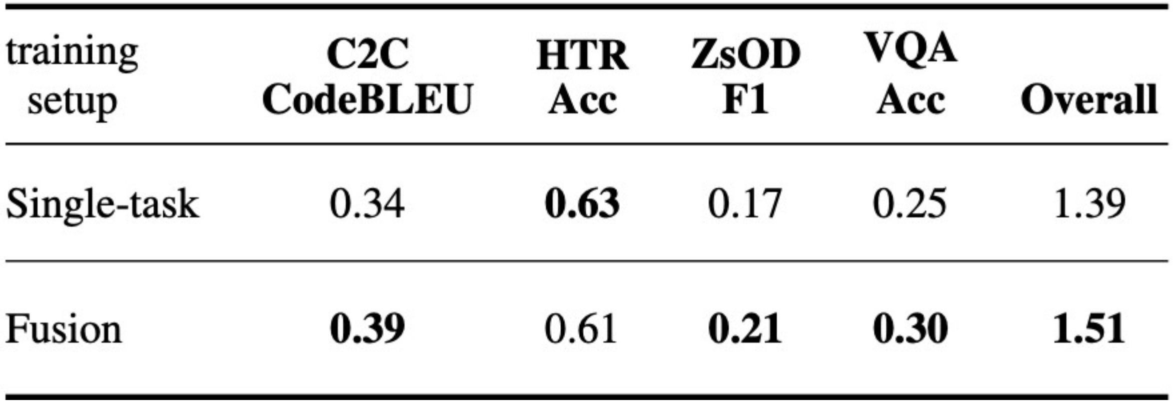
```java
class HelloWorld {
    public static void main(String[] args) {
        System.out.println("Hello, World!");
    }
}
```

```python
print('Hello, world!')
```

[1] Bakshandaeva, Daria, et al. "Many Heads but One Brain: an Overview of Fusion Brain Challenge on AI Journey 2021." 2021 (*Sber+AIRI*) - https://arxiv.org/abs/2111.10974

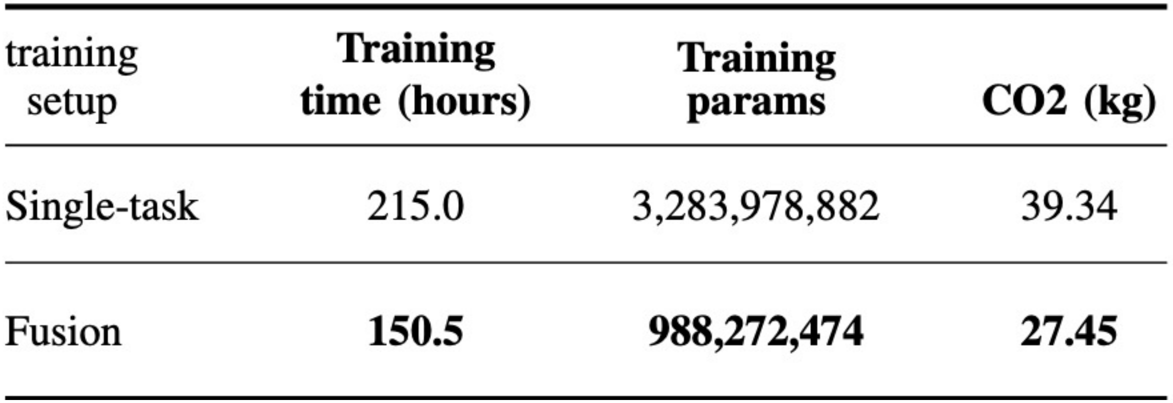
Fusion Brain approach[1]: results

Performance



training setup	C2C CodeBLEU	HTR Acc	ZsOD F1	VQA Acc	Overall
Single-task	0.34	**0.63**	0.17	0.25	1.39
Fusion	**0.39**	0.61	**0.21**	**0.30**	**1.51**

Efficiency



training setup	Training time (hours)	Training params	CO2 (kg)
Single-task	215.0	3,283,978,882	39.34
Fusion	**150.5**	**988,272,474**	**27.45**

For comparison:

	CO$_2$ emissions
Human Life	**5 ton**
Car with fuel	**57 ton**

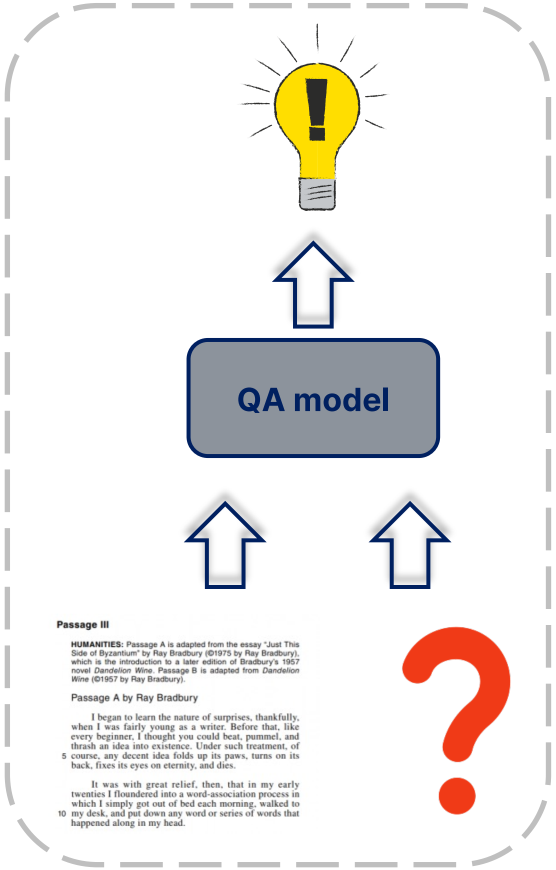
[1] Bakshandaeva, Daria, et al. "Many Heads but One Brain: an Overview of Fusion Brain Challenge on AI Journey 2021." 2021 (*Sber+AIRI*) - https://arxiv.org/abs/2111.10974

05

Retrieval-based models

Direction to add efficiency and explainability

Question Answering



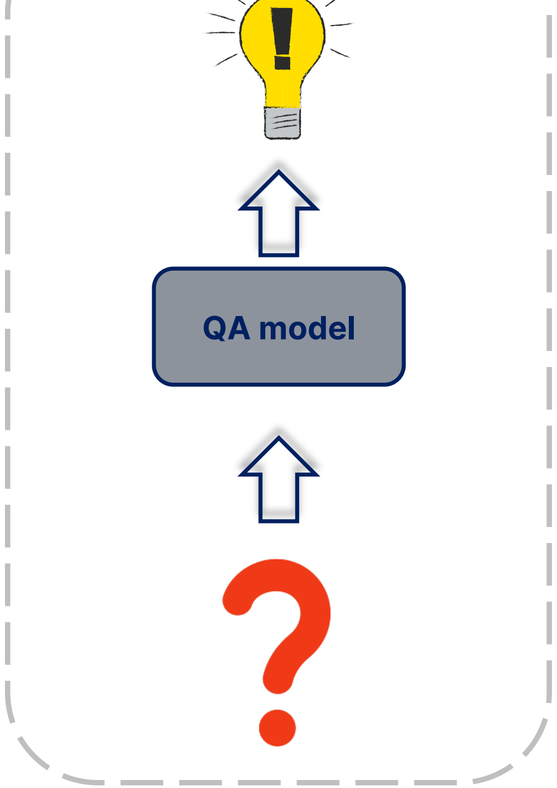
Passage III

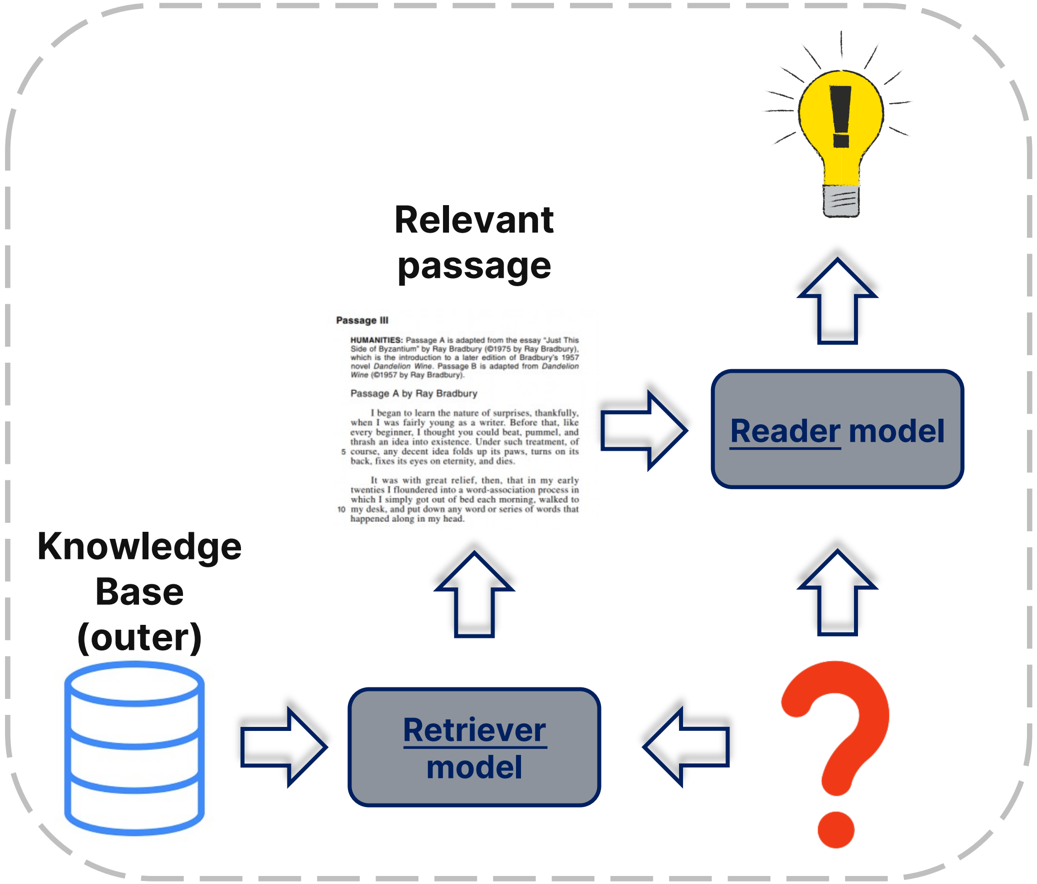
HUMANITIES: Passage A is adapted from the essay "Just This Side of Byzantium" by Ray Bradbury (©1975 by Ray Bradbury), which is the introduction to a later edition of Bradbury's 1957 novel *Dandelion Wine*. Passage B is adapted from *Dandelion Wine* (©1957 by Ray Bradbury).

Passage A by Ray Bradbury

I began to learn the nature of surprises, thankfully, when I was fairly young as a writer. Before that, like every beginner, I thought you could beat, pummel, and thrash an idea into existence. Under such treatment, of course, any decent idea folds up its paws, turns on its back, fixes its eyes on eternity, and dies.

It was with great relief, then, that in my early twenties I floundered into a word-association process in which I simply got out of bed each morning, walked to my desk, and put down any word or series of words that happened along in my head.





Passage III

HUMANITIES: Passage A is adapted from the essay "Just This Side of Byzantium" by Ray Bradbury (©1975 by Ray Bradbury), which is the introduction to a later edition of Bradbury's 1957 novel *Dandelion Wine*. Passage B is adapted from *Dandelion Wine* (©1957 by Ray Bradbury).

Passage A by Ray Bradbury

I began to learn the nature of surprises, thankfully, when I was fairly young as a writer. Before that, like every beginner, I thought you could beat, pummel, and thrash an idea into existence. Under such treatment, of course, any decent idea folds up its paws, turns on its back, fixes its eyes on eternity, and dies.

It was with great relief, then, that in my early twenties I floundered into a word-association process in which I simply got out of bed each morning, walked to my desk, and put down any word or series of words that happened along in my head.

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)

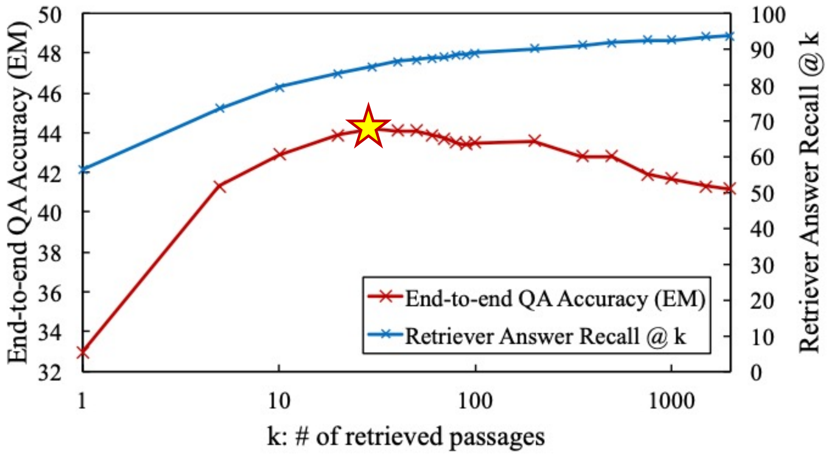
2-stage: first to retrieve the relevant model from outer text corpus, then extract knowledge from this passage
Realistic and scalable approach

[1] Roberts, Adam, Colin Raffel, and Noam Shazeer. "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" 2020 (*Google*)
[2] Brown, Tom B., et al. "Language models are few-shot learners." 2020 (*OpenAI*)

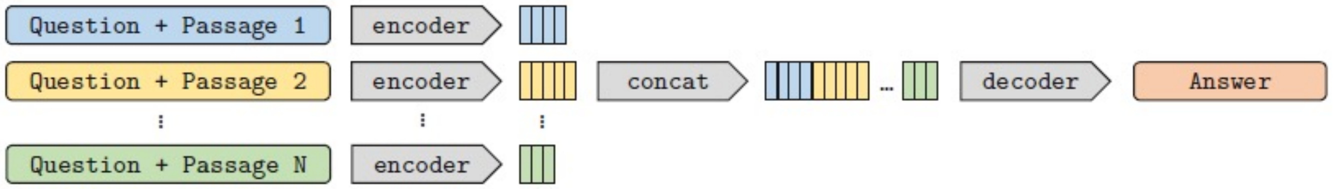
Retrieval-based (RB) modeling

WHY to decompose: 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.



How to extract information from multiple sources[2]



Main idea:

- Retriever **is not approx.** of Reader: having more data helps a little for the Reader, and drops quickly
- **Retriever** is a sort of **representational bottleneck**

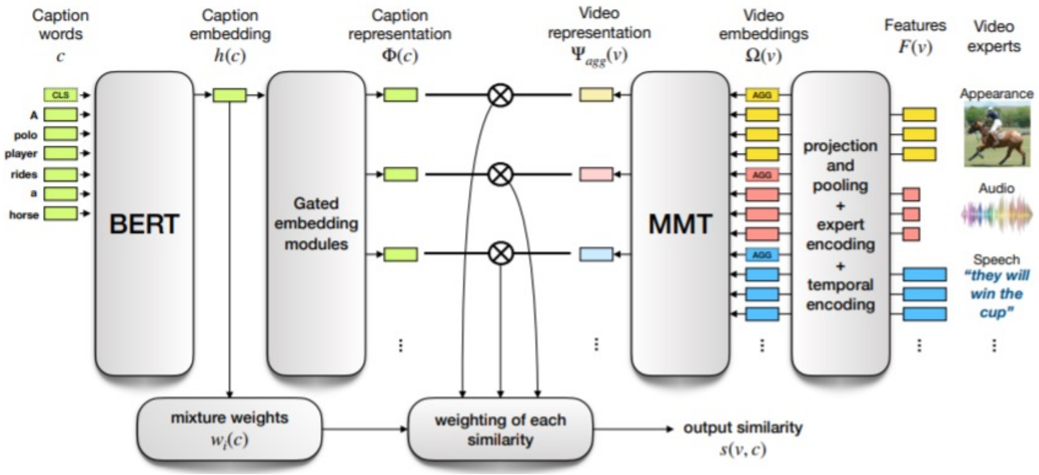
Main idea:

- **Retriever:** BERT-doc + BERT-query
- **Reader: 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 **k > 1 passages**

[1] Yang, Sohee, and Minjoon Seo. "Is Retriever Merely an Approximator of Reader?" 2020 (*NAVER Corp*)
[2] Izacard, Gautier, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." 2020 (*Facebook*)

Multi-modality and multi-task in RB

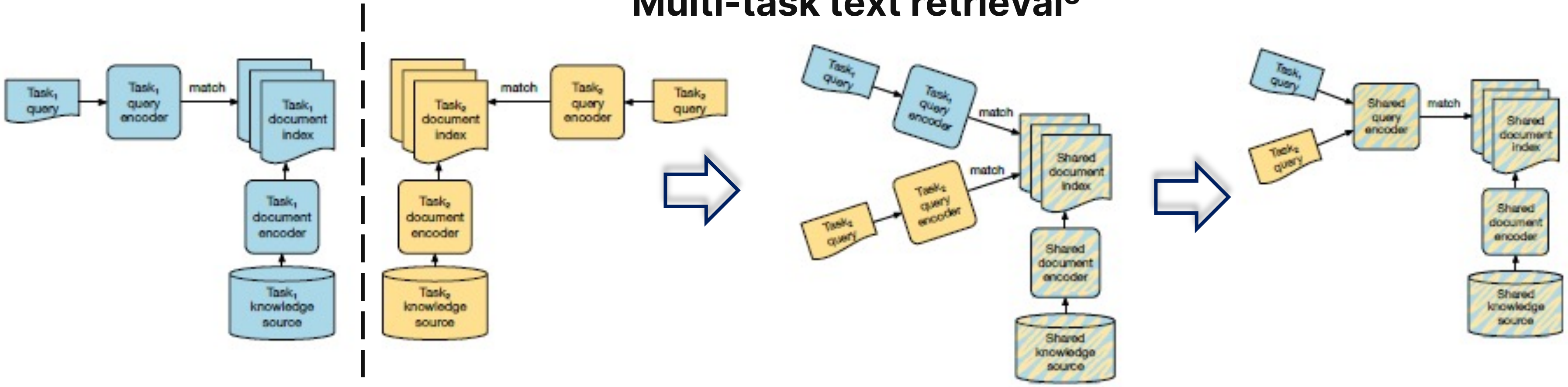
Multi-modality video retrieval[1,2]



Main idea:

- **Video** – as a **doc** in NLP RB, text query by BERT
- Multi-modality: **middle fusion** of non-query modalities + **late fusion** with text query
- For different NLP tasks the **single retriever is beneficial**
- But the **training** of retriever should be done on **all datasets combined**
- **Retriever: BERT**-based; **Reader**/downstream: **BART**-based

Multi-task text retrieval[3]



[1] Gabeur, Valentin, et al. "Multi-modal transformer for video retrieval." 2020 (*Google*)
[2] Dzabraev, Maksim, et al. "Mdmmt: Multidomain multimodal transformer for video retrieval." 2021 (*Huawei*)
[3] Maillard, Jean, et al. "Multi-task retrieval for knowledge-intensive tasks." 2021 (*Facebook*)

06

Open Questions

Open Questions

1. **Effectiveness**

   Current trend: usage of *LARGE* pre-trained models

   Q: How to *decrease* the *resource utilization* (while training as well as on inference)?

2. **Universality**

   Q_1: How to add the new modality *agnostically* (with minimal architectural changes)?

   Q_2: How to add the new task *agnostically* (without full retraining)?

   Q_3: What tasks could and what tasks should not be combined?

AIRI

airi.net