Duel-based Deep Learning system for solving IQ tests

Paulina Tomaszewska,
Adam Żychowski,
Jacek Mańdziuk

Warsaw University of Technology

Conference: Artificial Intelligence and Statistics (AISTATS) 2022

	Publication	h5-index	h5-median
1.	International Conference on Learning Representations	253	470
2.	Neural Information Processing Systems	245	422
3.	International Conference on Machine Learning	204	370
4.	AAAI Conference on Artificial Intelligence	157	240
5.	IEEE Transactions On Systems, Man And Cybernetics Part B, Cybernetics	127	172
6.	IEEE Transactions on Neural Networks and Learning Systems	119	171
7.	Neurocomputing	119	164
8.	Expert Systems with Applications	118	164
9.	International Joint Conference on Artificial Intelligence (IJCAI)	105	174
10.	Applied Soft Computing	103	133
11.	Journal of Machine Learning Research	96	165
12.	IEEE Transactions on Fuzzy Systems	96	128
13.	Knowledge-Based Systems	96	127
14.	Neural Computing and Applications	83	115
15.	Neural Networks	72	105
16.	International Conference on Artificial Intelligence and Statistics	68	101
17.	Engineering Applications of Artificial Intelligence	65	93
18.	Robotics and Autonomous Systems	58	91
19.	Conference on Learning Theory (COLT)	57	101
20.	International Joint Conference on Neural Networks	57	84

https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence

- double-blind submissions (8 pages excluding references)

Review process:

- in order to submit paper, it was **obligatory** to nominate at least one author as a reviewer
- reviewer have access to a list of abstacts of submitted papers and specify the degree in which he is interested in reviewing particular paper

Review form

Summary and contributions

Strengths

Weaknesses

Correctness

Clarity

Relation to prior work

Additional Comments

Reproducibility

Assumptions and limitations

Societal impact

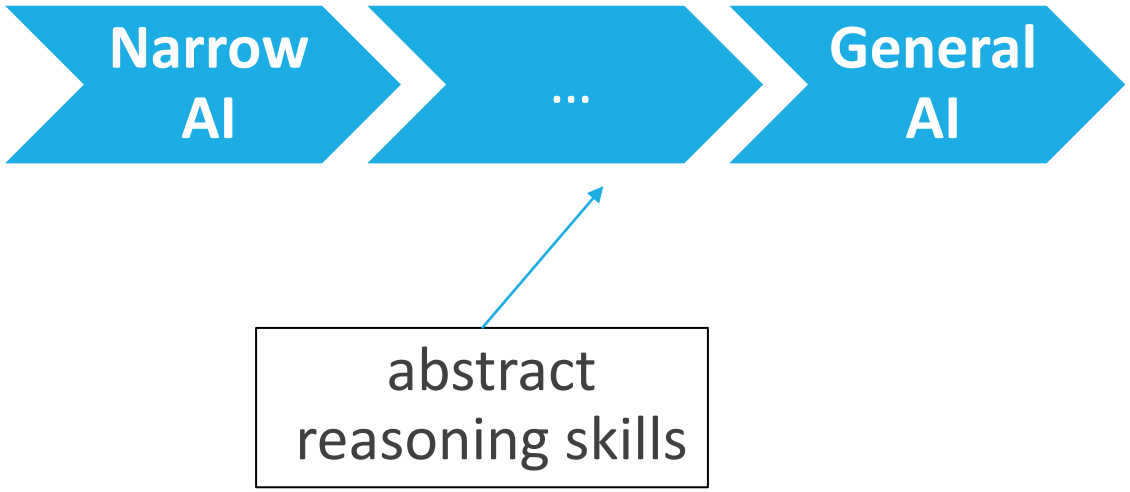
Code release

Score

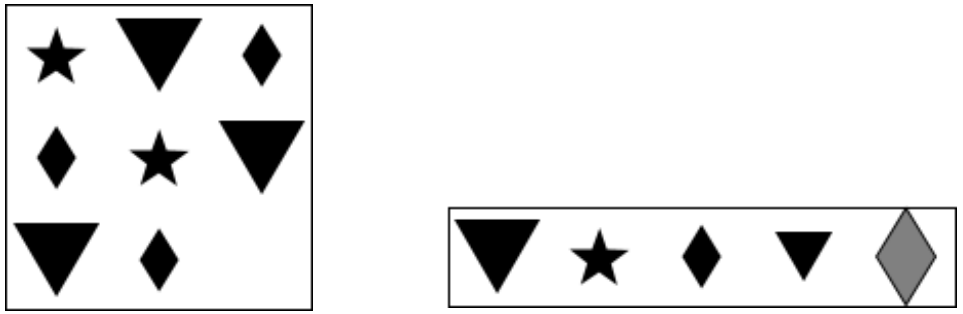
Confidence score

Ethical concerns

Motivation



How to assess abstract reasoning skills?
- IQ tests in a form of Raven's Progressive Matrices (RPMs)

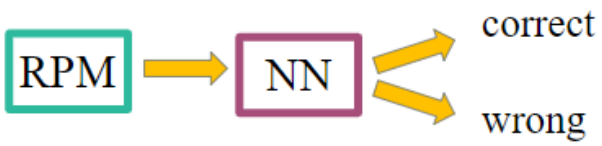


Example of an RPM task (**left**) with the set of *K* candidate answers (**right**). The correct answer is the second from the left.

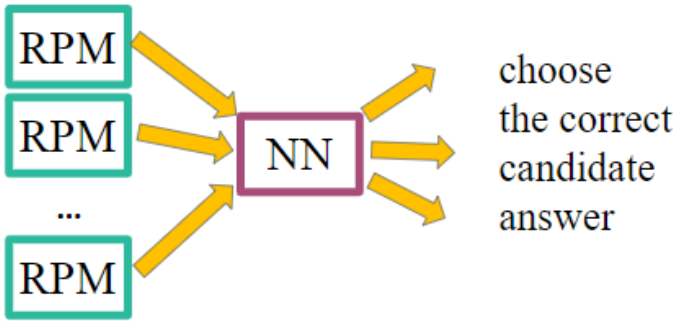
Approaches

Possible architectures
(two extremes):

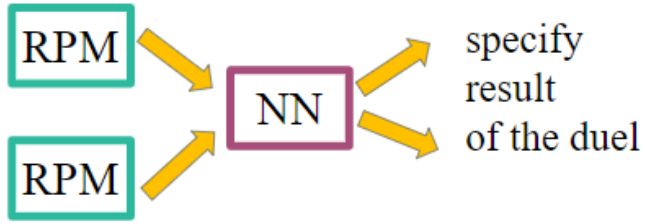
- 1 - input



- *K* - input



Alternative (trade-off) approach
(**tournament with duels**):

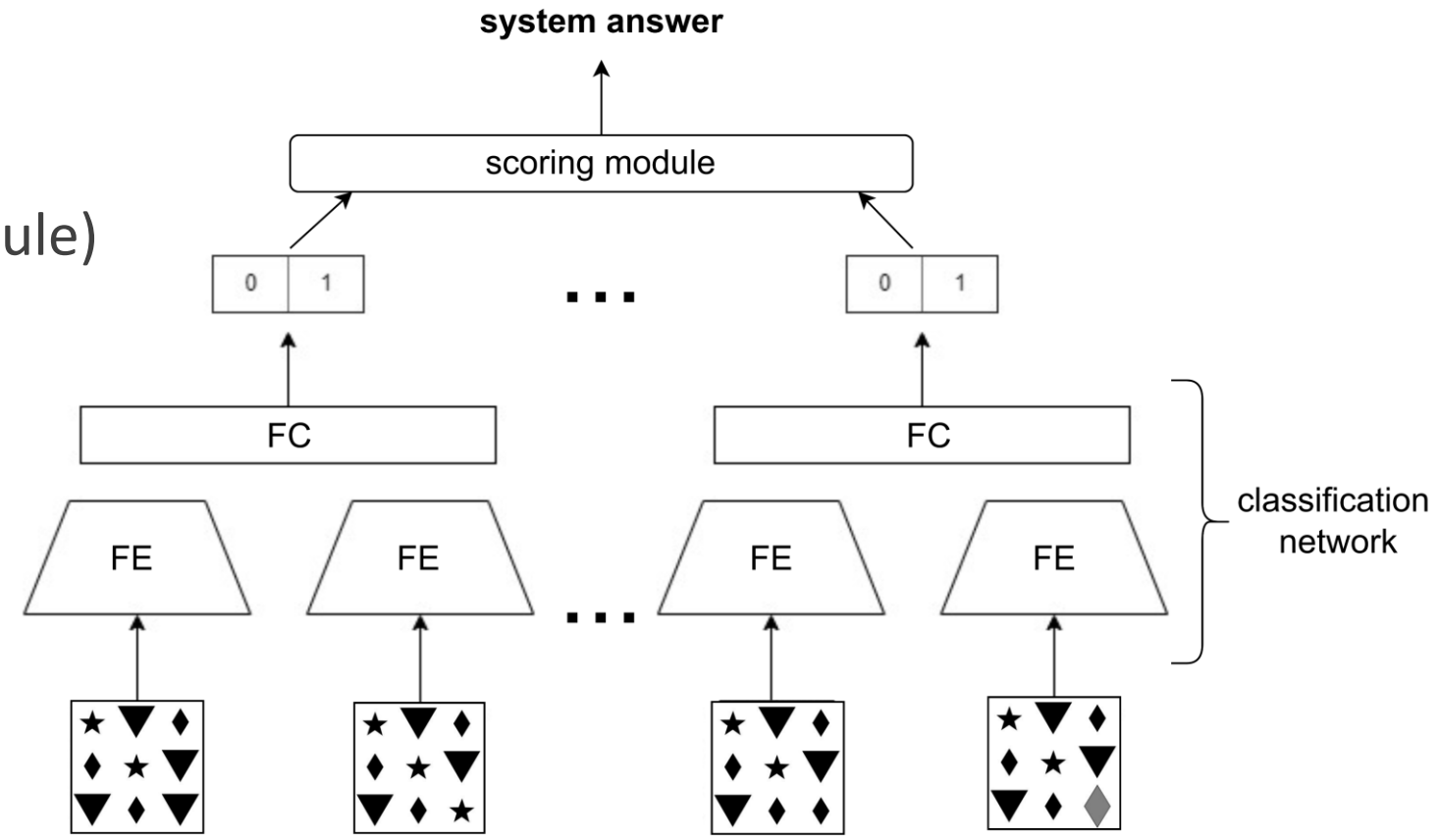


Assumption:

- *2*-class classification should be easier than *K*-class classification

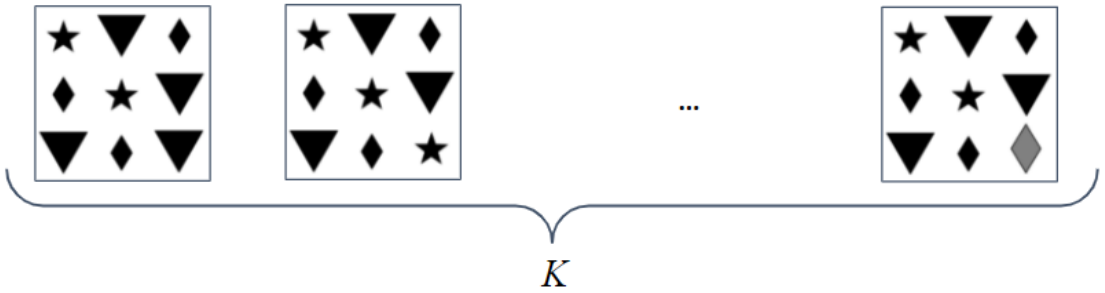
Proposed system: Duel-IQ

1. preprocessing
2. Deep Learning model
3. postprocessing (scoring module)

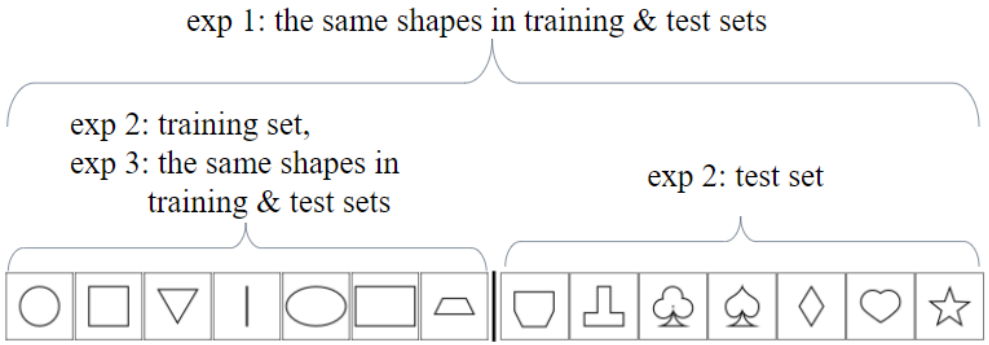


Preprocessing

filled-in RPMs



dataset



Pairing schemes for the duels:

- ***winner_in***

within the input pair, there is always a correct answer

- ***any_in***

there is no such a requirement as in *winner_in*

(a draw in the duel is possible)

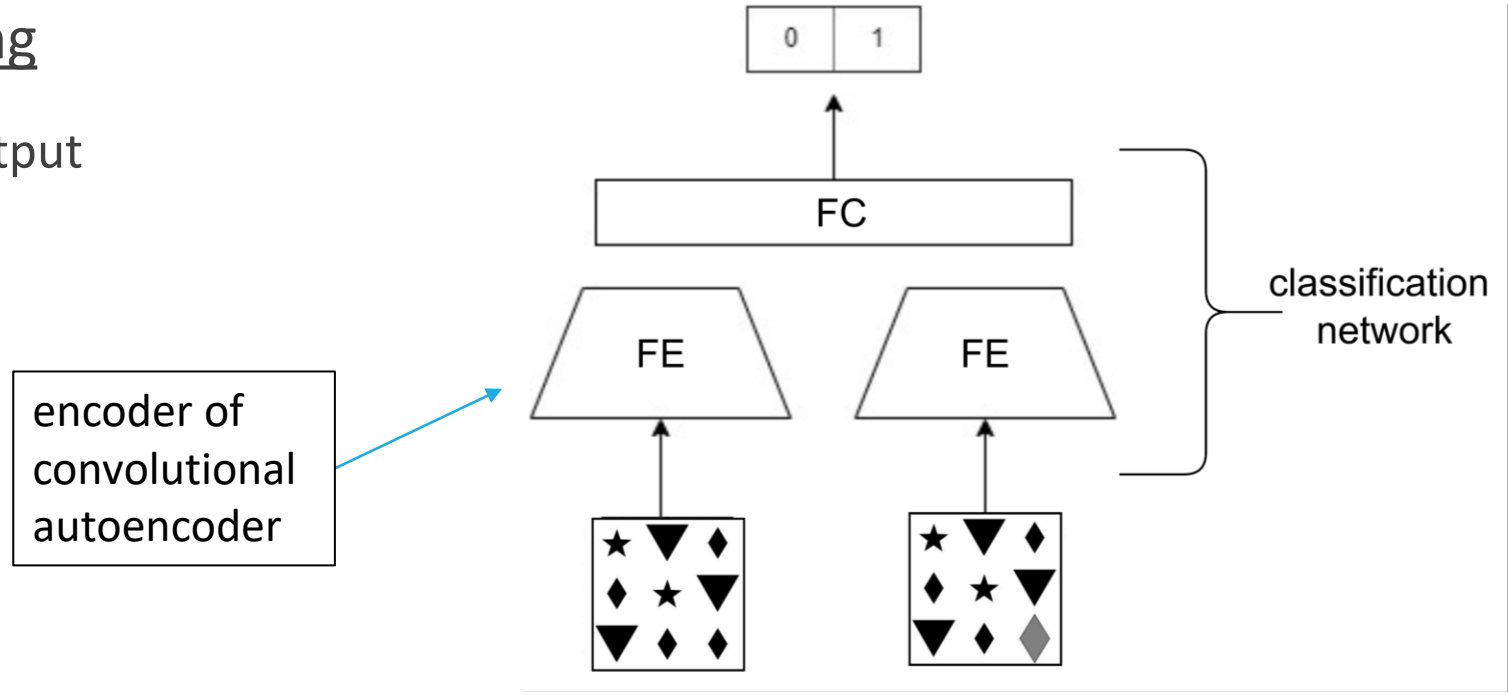
Deep Learning model

- with/without auxiliary training

in auxiliary training, the target is to output

a dissmilarity score between the given

candidate answers



model in *winner_in* variant

*classification for brevity

Scoring module

- aggregation of the results of duels
- two scoring functions:

  - ***winner frequency***

  the final system output is the answer that won the biggest number of duels

  - ***probability sum***

  the winner is the answer with the highest overall sum of probabilities of winning the duels

Results

ALGORITHM	EXP 1	EXP 2	EXP 3
Duel-IQ	**82.8 ± 6.2** (**88.8**)	49.1 ± 1.8 (54.9)	**80.0 ± 3.2** (**85.3**)
Duel-IQ (5-class version)	72.2 ± 5.4 (75.6)	49.2 ± 1.1 (50.8)	71.5 ± 3.2 (74.1)
DeepIQ	73.0 ± 3.7 (78.3)	**61.7 ± 3.9** (**67.4**)	68.8 ± 2.4 (72.5)
WReN	54.1 ± 1.1 (56.3)	34.1 ± 1.2 (36.0)	49.0 ± 1.5 (51.1)
IQ of NN (2-class version)	67.3 ± 20.1 (90.3)	49.1 ± 10.2 (72.8)	68.3 ± 11.2 (89.8)
IQ of NN (5-class version)	55.8 ± 11.5 (76.3)	44.3 ± 5.7 (54.8)	55.1 ± 6.7 (69.3)

Human performance (exp 1)

77.3% ± 5.6%

1. When the same data distribution in training and test sets:
   Duel-IQ is superior to other methods of similar complexity
2. In out-of-distribution case (not the main interest in our work):
   DeepIQ outperforms Duel-IQ
3. Results of Duel-IQ are on par with human performance