Analiza relacji w obrazach z wykorzystaniem sieci neuronowych

Abstract relational reasoning in neural networks

Mikołaj Małkiński

21 kwietnia 2020

Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

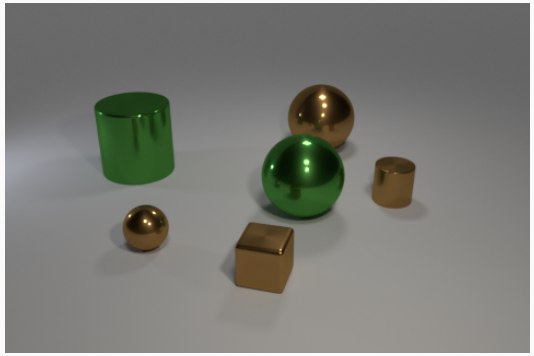
Spis treści

Visual Question Answering

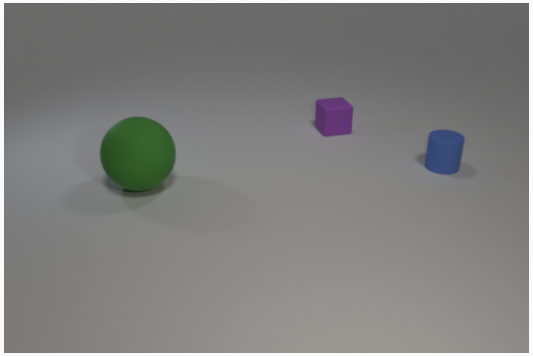
Visual Question Answering

- Połączenie prostych zagadnień przetwarzania języka naturalnego (NLP) z wizją komputerową (CV)
- Komponent NLP jest zazwyczaj prosty, reprezentuje pytania za pomocą one-hot encoding lub sieci rekurencyjnej LSTM/GRU
- Główną trudnością jest powiązanie obiektów między pytaniem a obrazem
- Pytania dotyczą porównań, zliczania, rozpoznania atrybutów wizualnych i operacji logicznych

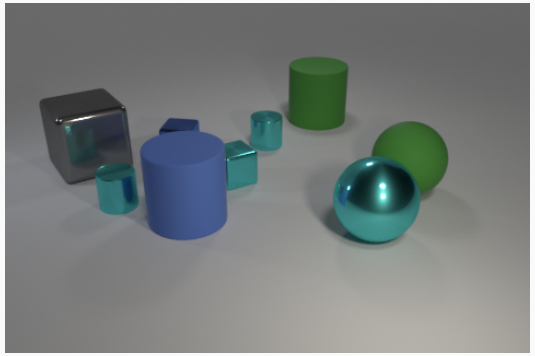
Visual Question Answering



Q: What is the material of the tiny object to the right of the brown shiny ball behind the tiny shiny cylinder?
A: metal



Q: The matte thing that is both in front of the purple cube and to the left of the blue rubber cylinder is what color?
A: green



Q: Are there fewer metallic objects that are on the left side of the large cube than cylinders to the left of the cyan shiny block?
A: yes

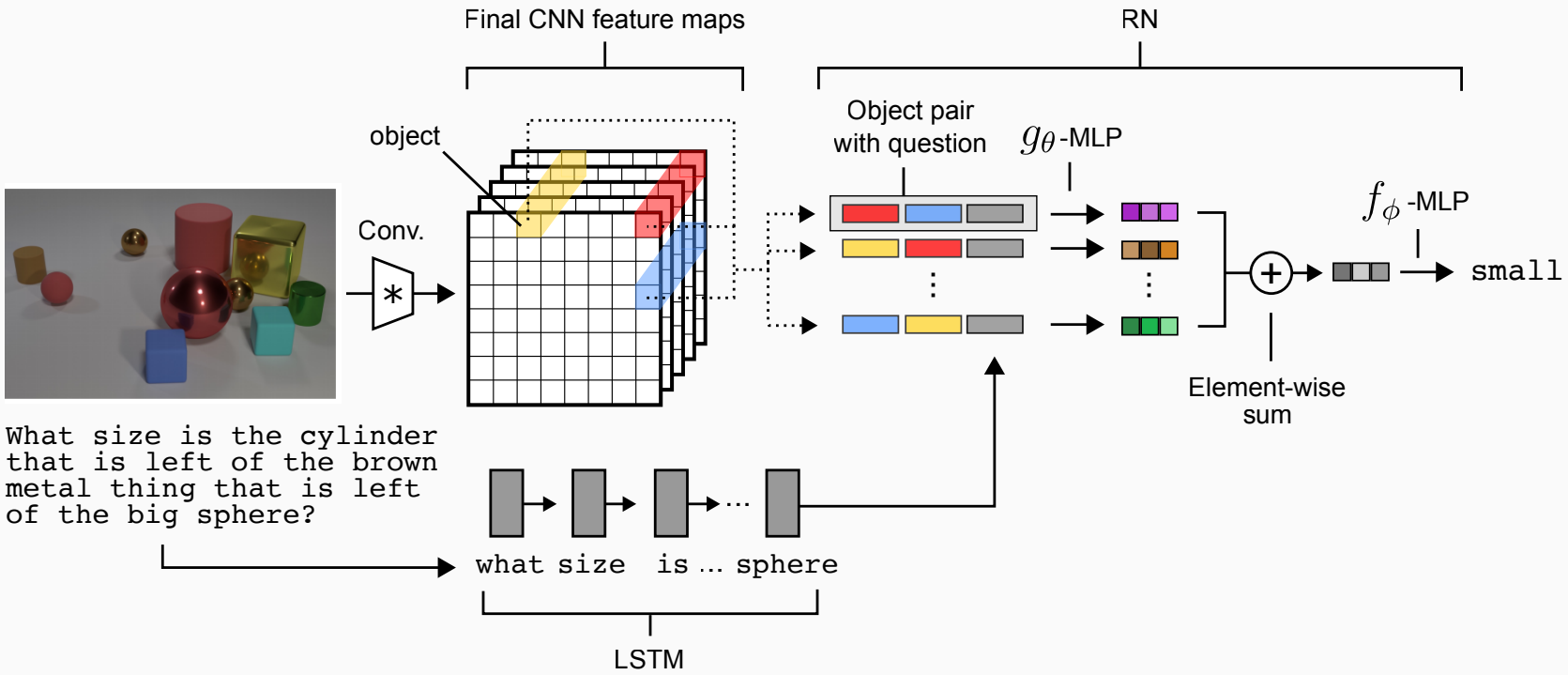
Rysunek 1: Przykładowe obrazy, pytania i odpowiedzi z zbioru CLEVR [4].

Sieć relacyjna (Relation Network) [5]

$$RN(O) = f_\phi(\sum_{i,j} g_\theta(o_i, o_j)) \tag{1}$$

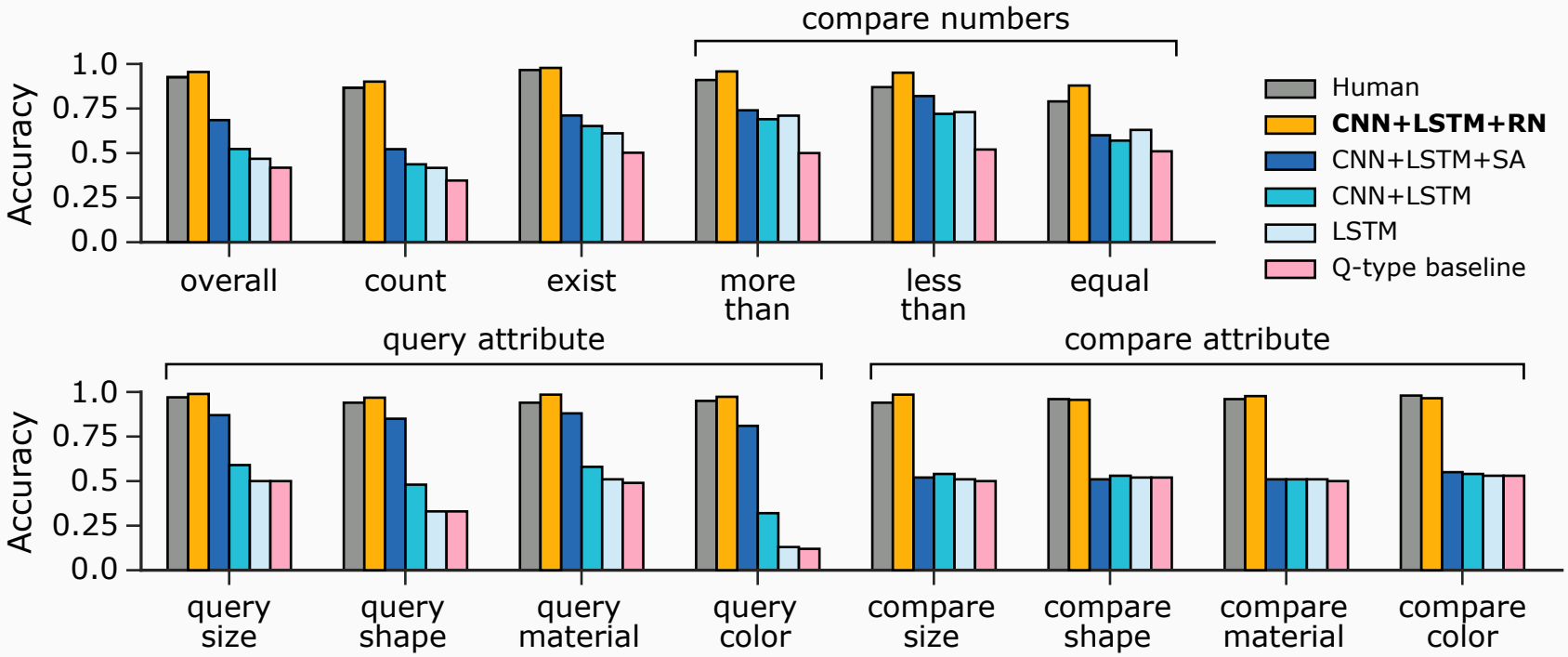
- Prosty moduł który może być łatwo dodany do istniejących architektur
- Działa na zbiorze obiektów, niezależnie od ich reprezentacji
- Uwzględnia relację pomiędzy wszystkimi parami obiektów
- Wynik nie zależy od kolejności obiektów w zbiorze

Sieć relacyjna



Rysunek 2: Architektura wykorzystująca sieć relacyjną dla zbioru CLEVR.

Sieć relacyjna



Rysunek 3: Wyniki sieci relacyjnej na zbiorze CLEVR.

Sieć relacyjna

Inne problemy w których sieć relacyjna osiąga wysokie wyniki:

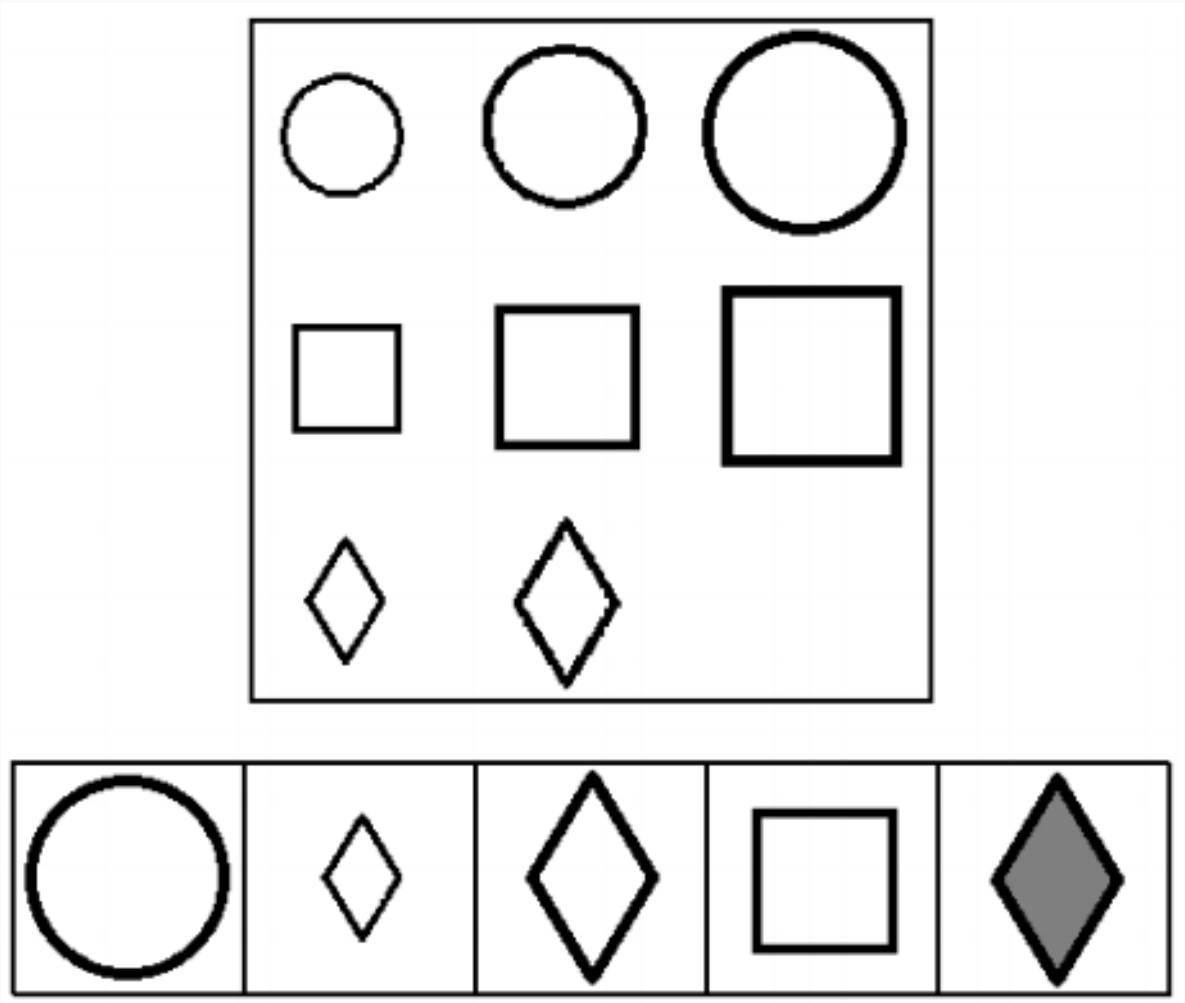
- Sort-of-CLEVR [5] - wizualnie prostsza, dwuwymiarowa, wersja zbioru CLEVR która rozdziela pytania między relacyjne a nierelacyjne
- bAbI [6] - w pełni tekstowy zbiór pytań i odpowiedzi
- Dynamic physical systems [5] - zbiór przedstawiający poruszające się obiekty z ukrytyi powiązaniami

Progresywne Matryce Ravena

Progresywne Matryce Ravena (Raven's Progressive Matrices)

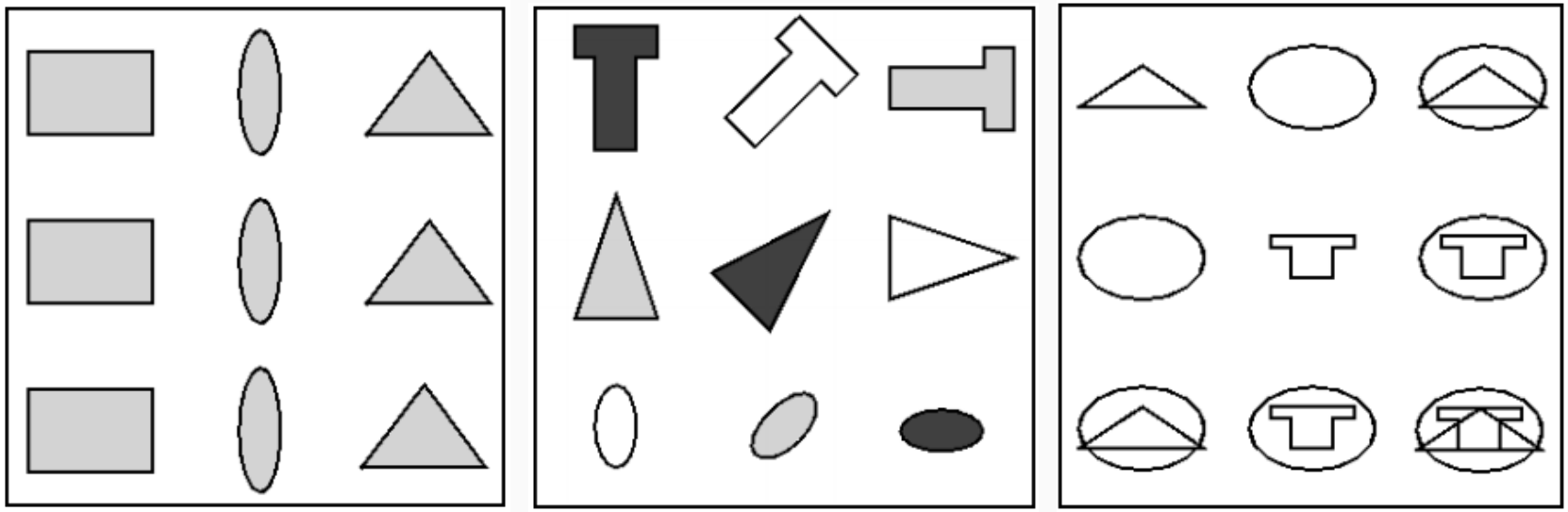
- Prosta wizualna reprezentacja, składająca się z figur o różnym kształcie, kolorze, rotacji i pozycji względem reszty
- Problem polega na zidentyfikowaniu relacji (jednej lub wielu) rządzącej atrybutami figur
- Często używane jako wyznacznik ludzkiej inteligencji

Progresywne Matryce Ravena



Rysunek 4: Przykład prostego RPM - dobrą odpowiedzią jest trzecia figura [1].

Progresywne Matryce Ravena



Rysunek 5: Przykłady uzupełnionych matryc z różnym poziomem trudności [1].

Zbiór danych PGM

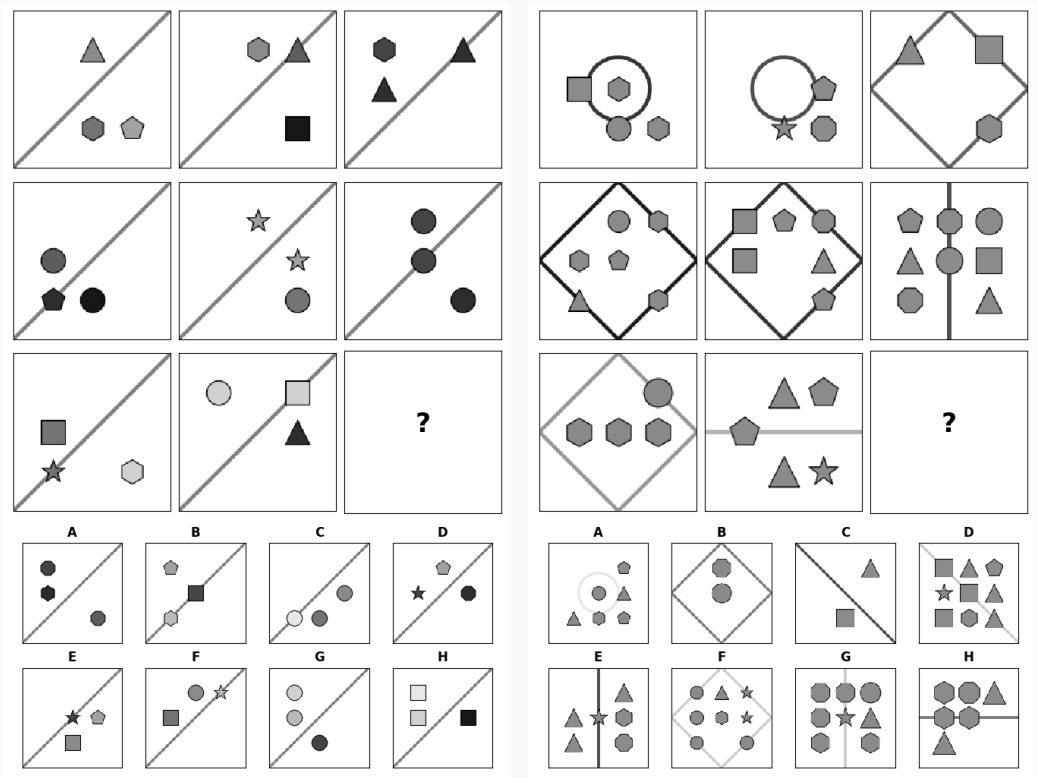
Motywacja

Czy sieci neuronowe są w stanie rozwiązać abstrakcyjne wizualne problemy wymagające rozumowania o relacjach? W jaki sposób wytrenować taką sieć?

Zbiór danych PGM

- Zbiór złożony z matryc RPM z powiązanymi strukturami:
  $\mathcal{S} = \{[r, o, a] | r \in \mathcal{R}, o \in \mathcal{O}, a \in \mathcal{A}\}$
- Relacje:
  $\mathcal{R} = \{\texttt{progression}, \texttt{XOR}, \texttt{OR}, \texttt{AND}, \texttt{consistent union}\}$
- Obiekty: $\mathcal{O} = \{\texttt{shape}, \texttt{line}\}$
- Atrybuty: $\mathcal{A} = \{\texttt{size}, \texttt{type}, \texttt{color}, \texttt{position}, \texttt{number}\}$
- $1 \leq |\mathcal{S}| \leq 4$
- Rozpraszające losowe atrybuty które nie należą do żadnej struktury utrudniają znalezienie dobrego rozwiązania

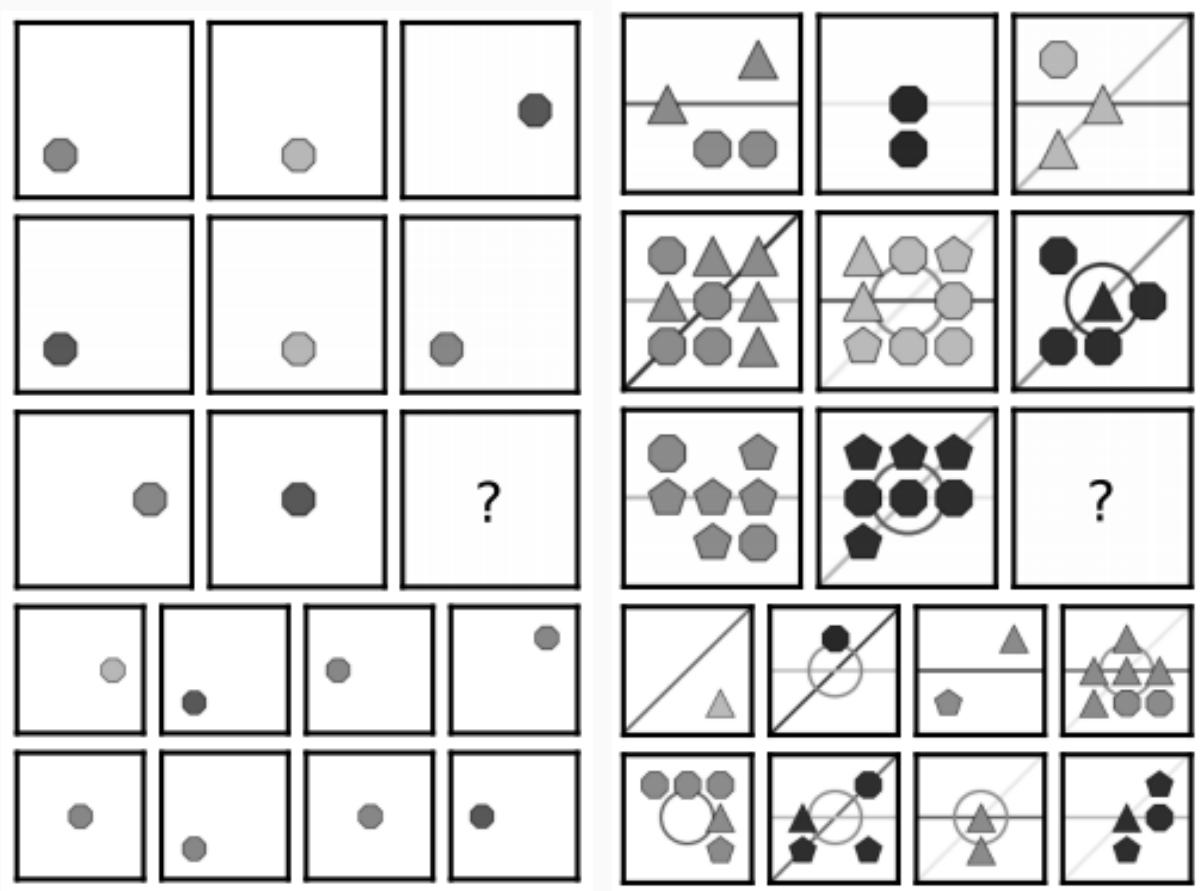
Zbiór danych PGM



(a) **(b)**

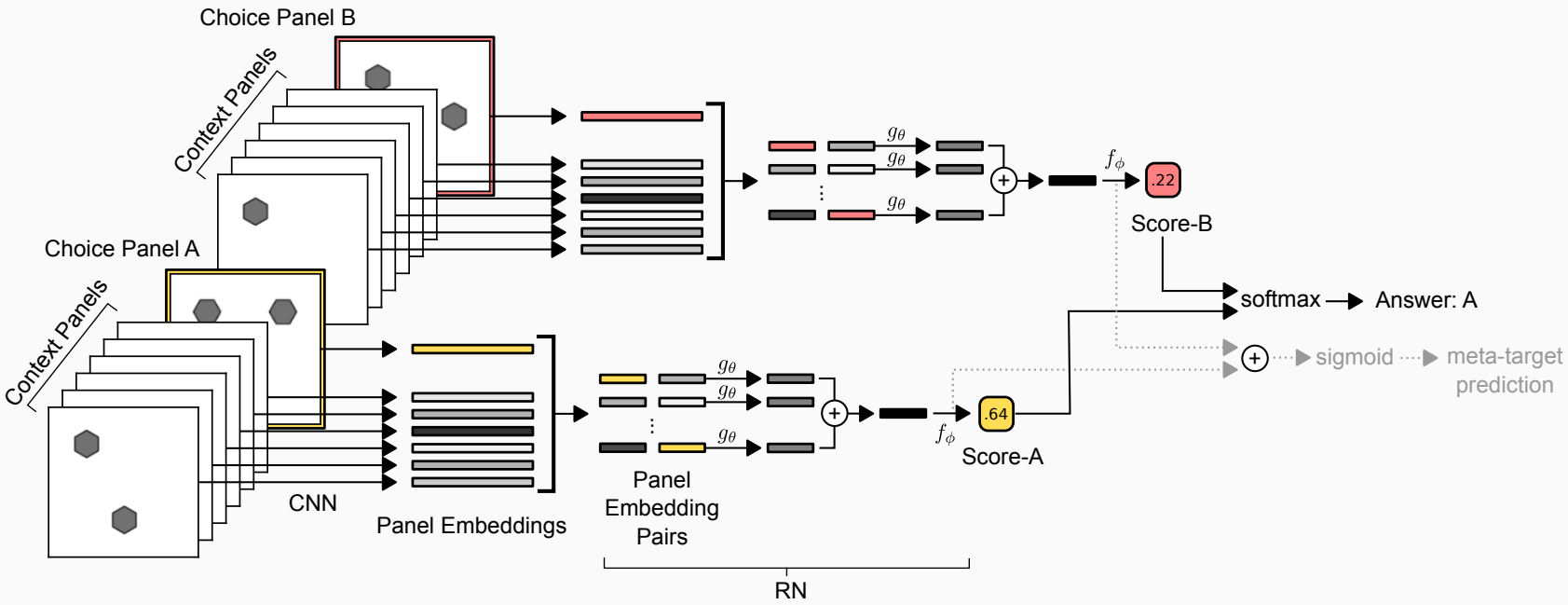
Rysunek 6: Przykłady z zbioru PGM: a) zawiera jedną relację [`AND, shape, type`], poprawna odpowiedź to H; b) zawiera 2 relacje [`progression, shape, size`] i [`progression, line, color`], poprawna odpowiedź to G.

Zbiór danych PGM



Rysunek 7: Wpływ rozpraszających atrybutów - w obu przypadkach relacja to [`consistent union, shape, color`]

Wild Relation Network [2]



Rysunek 8: Model WReN wykorzystujący sieć relacyjną.

Wild Relation Network

Rysunek 9: Wyniki modeli na neutralnej konfiguracji generalizacyjnej

Model	Test (%)
WReN	62.6
Wild-ResNet	48.0
ResNet-50	42.0
LSTM	35.8
CNN + MLP	33.0
Blind ResNet	22.4

Zbiór danych PGM

Konfiguracje generalizacyjne:

- neutralna - zbiór treningowy i testowy może zawierać każdą trójkę $[r, o, a]$
- interpolacja - $a \in \{\texttt{color}, \texttt{size}\}$, zbiór treningowy zawiera nieparzyste wartości V_a a testowy parzyste
- ekstrapolacja - $a \in \{\texttt{color}, \texttt{size}\}$, zbiór treningowy zawiera dolne wartości V_a a testowy górne
- ukryte atrybuty `shape-color`
- ukryte atrybuty `line-type`
- ukryte trójki
- ...

Trening pomocniczy

Wzbogacenie funkcji kosztu o wartość związaną z przewidywaniem relacji (struktur) w danym problemie:

$$\mathcal{L}_{total} = \mathcal{L}_{target} + \beta \mathcal{L}_{meta-target} \tag{2}$$

Wild Relation Network

Rysunek 10: Wyniki modelu WReN na wszystkich konfiguracjach generalizacyjnych

	$\beta = 0$		$\beta = 10$	
Regime	Val. (%)	Test (%)	Val. (%)	Test (%)
Neutral	63.0	62.6	77.2	76.9
Interpolation	79.0	64.4	92.3	67.4
H.O. Attribute Pairs	46.7	27.2	73.4	51.7
H.O. Triple Pairs	63.9	41.9	74.5	56.3
H.O. Triples	63.4	19.0	80.0	20.1
H.O. line-type	59.5	14.4	78.1	16.4
H.O. shape-colour	59.1	12.5	85.2	13.0
Extrapolation	69.3	17.2	93.6	15.5

Wild Relation Network - wnioski

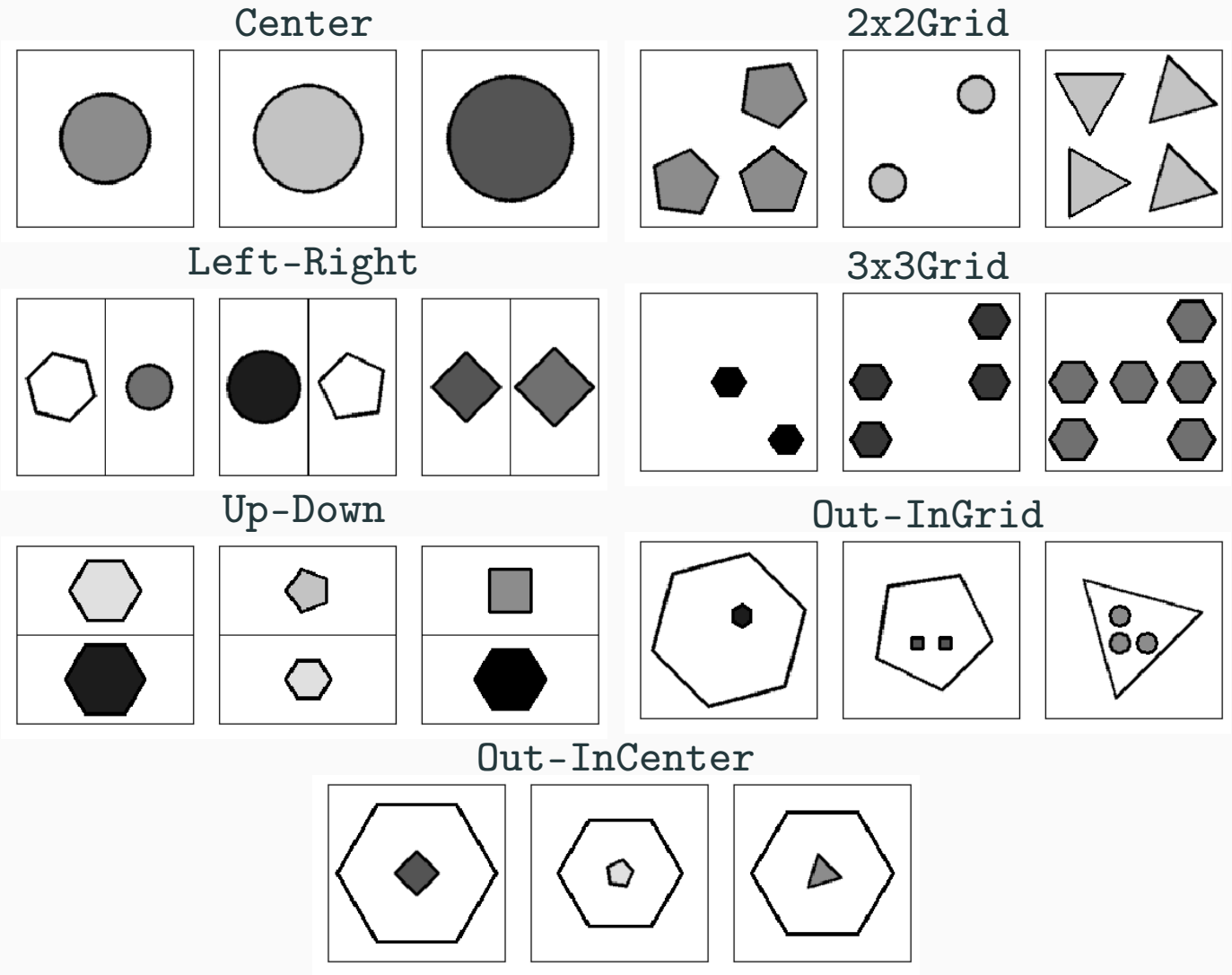
- Sieci neuronowe są w stanie rozumować o abstrakcyjnych relacjach i tworzyć reprezentacje logicznych i matematycznych operacji na podstawie surowych pixeli obrazu
- Architektura modelu ma duży wpływ na wyniki
- Wyniki mogą być ulepszone za pomocą dodatkowego treningu, który pomaga sieci budować bardziej trafne reprezentacje

Zbiór danych RAVEN

Motywacja

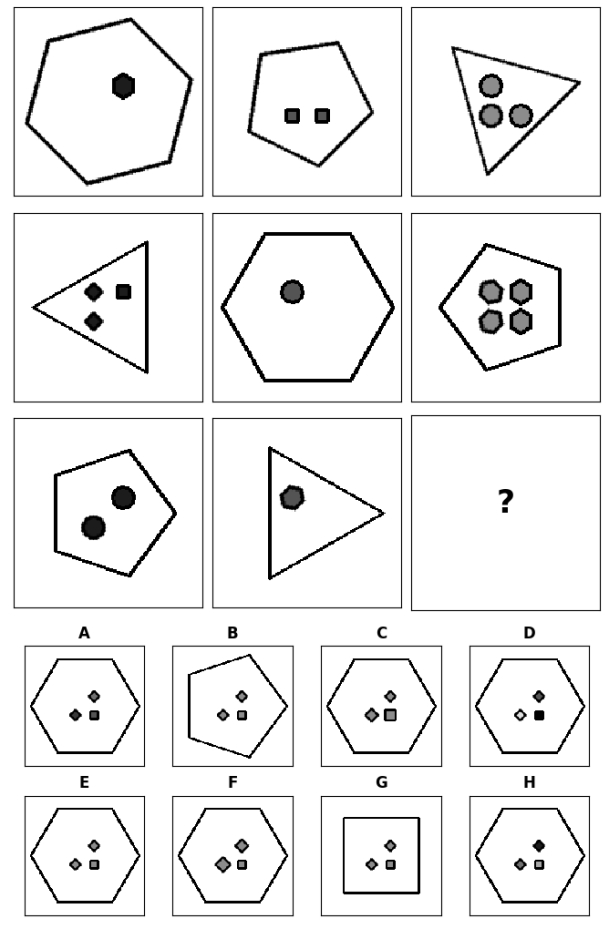
- Użycie hierarchicznych reprezentacji matryc
- Dodanie annotacji strukturalnych
- Ocena wyników ludzkich

Zbiór danych RAVEN [7]



Rysunek 11: Konfiguracje matryc z zbioru RAVEN.

Zbiór danych RAVEN



Rysunek 12: Przykład z zbioru RAVEN, poprawna odpowiedź to E.

Wyniki na zbiorze RAVEN

Rysunek 13: Wyniki baseline'owych modeli na zbiorze RAVEN.

Method	Acc (%)
LSTM	13.07
WReN	14.69
CNN	36.97
ResNet	53.43
LSTM+DRT	13.96
WReN+DRT	15.02
CNN+DRT	39.42
ResNet+DRT	59.56
Human	84.41
Solver	100

Wyniki na zbiorze RAVEN

Rysunek 14: Szczegółowe wyniki baseline'owych modeli na zbiorze RAVEN.

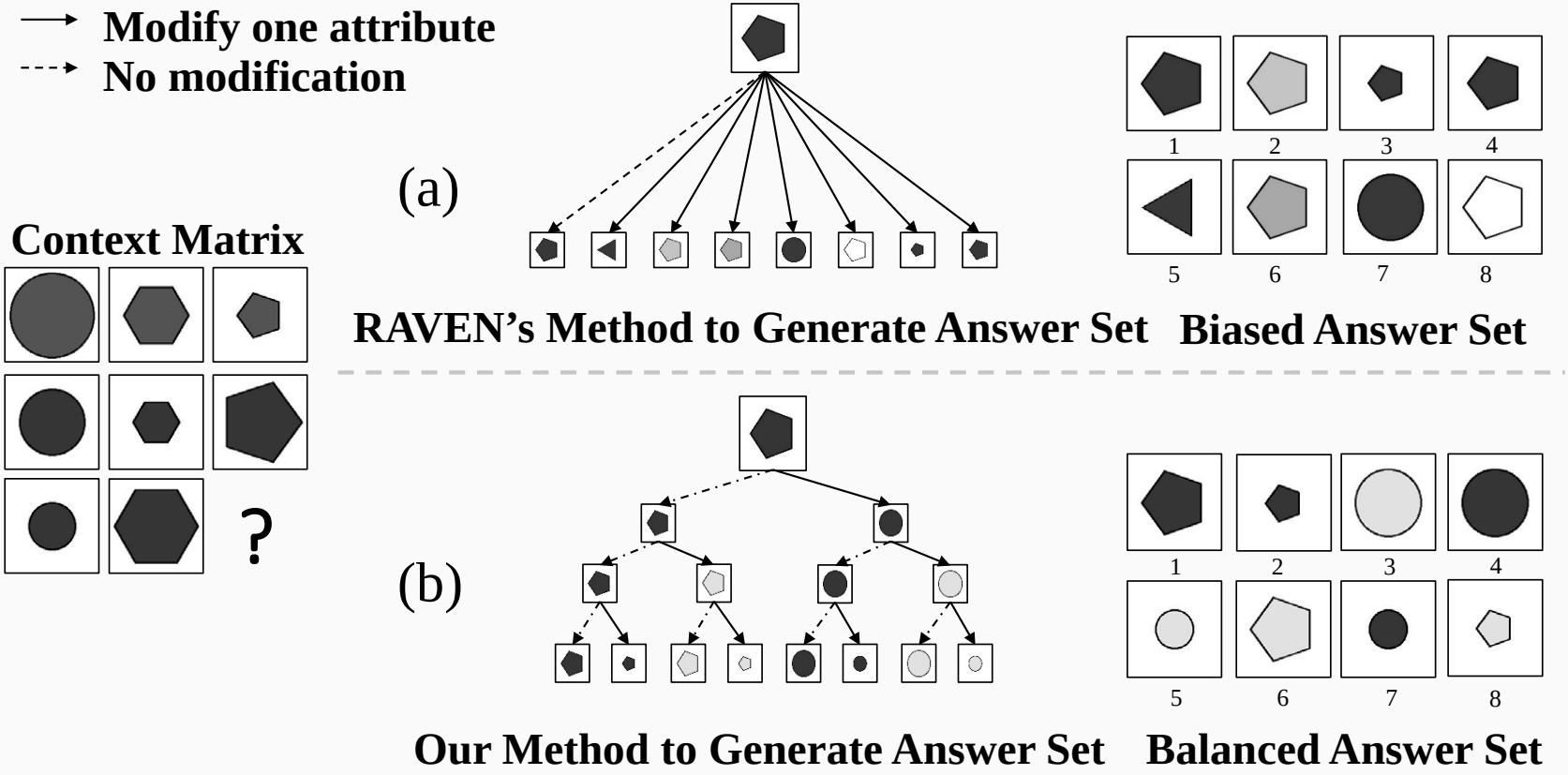
Method	Acc (%)	Center (%)	2x2Grid (%)	3x3Grid (%)	L-R (%)	U-D (%)	O-IC (%)	O-IG (%)
LSTM	13.07	13.19	14.13	13.69	12.84	12.35	12.15	12.99
WReN	14.69	13.09	28.62	28.27	7.49	6.34	8.38	10.56
CNN	36.97	33.58	30.30	33.53	39.43	41.26	43.20	37.54
ResNet	53.43	52.82	41.86	44.29	58.77	60.16	63.19	53.12
LSTM+DRT	13.96	14.29	15.08	14.09	13.79	13.24	13.99	13.29
WReN+DRT	15.02	15.38	23.26	29.51	6.99	8.43	8.93	12.35
CNN+DRT	39.42	37.30	30.06	34.57	45.49	45.54	45.93	37.54
ResNet+DRT	59.56	58.08	46.53	50.40	65.82	67.11	69.09	60.11
Human	84.41	95.45	81.82	79.55	86.36	81.81	86.36	81.81
Solver	100	100	100	100	100	100	100	100

Problem zbioru RAVEN

Rysunek 15: Porównanie wyników baseline'owych modeli między zbiorem RAVEN a Balanced-RAVEN [3], wykorzystując prostą augmentację danych.

Model	RAVEN	Balanced-RAVEN
ResNet	89.2%	40.3%
Context-blind ResNet	90.1%	12.5%

Zbiór danych Balanced-RAVEN [3]



Rysunek 16: Porównanie sposobów generowania odpowiedzi.

Hierarchical Rule Induction Network (HriNet) [3]

Model próbuje odwzorować ludzkie podejście do rozwiązania problemu:

1. identyfikacja podstawowych atrybutów (typ, kolor, rozmiar)
2. porównanie paneli w danym wierszu w celu znalezienia możliwych relacji
3. porównanie paneli w dwóch wierszach w celu znalezienia wspólnych relacji
4. uzupełnienie trzeciego wiersza kandydatem odpowiedzi i znalezienie wspólnych cech z już istniejącymi wierszami

Model HriNet

Na podstawie dwóch wierszy M_i i M_j, model poszukuje relacji z poziomu różnych hierarchii:

- *cell-wise hierarchy* - reprezentacje osobnych paneli w wierszu: $x_{ij} = \mathbb{E}_{cell}(m_{ij})$
- *individual-wise hierarchy* - reprezentacja całego wiersza: $y_i = \mathbb{E}_{ind}(M_i)$
- *ecological hierarchy* - reprezentacja dwóch wierszy: $z_{ij} = \mathbb{E}_{eco}([M_i, M_j])$

Każda hierarchia jest reprezentowana przez osobną sieć konwolucyjną z różną ilością kanałów wejściowych.

Model HriNet

Reprezentacje różnych hierarchii koncentrują się na odrębnych szczegółach paneli. Agregacja hierarchii:

- na poziomie *cell-wise*: $r_i^{(1)} = \varphi_1(x_{i1}, x_{i2}, x_{i3})$
- na poziomie *individual-wise*: $r_{ij}^{(2)} = \varphi_2(r_i^{(1)}, y_i, r_j^{(1)}, y_j)$
- na poziomie *ecological*: $r_{ij}^{(3)} = \varphi_3(r_{ij}^{(2)}, z_{ij})$

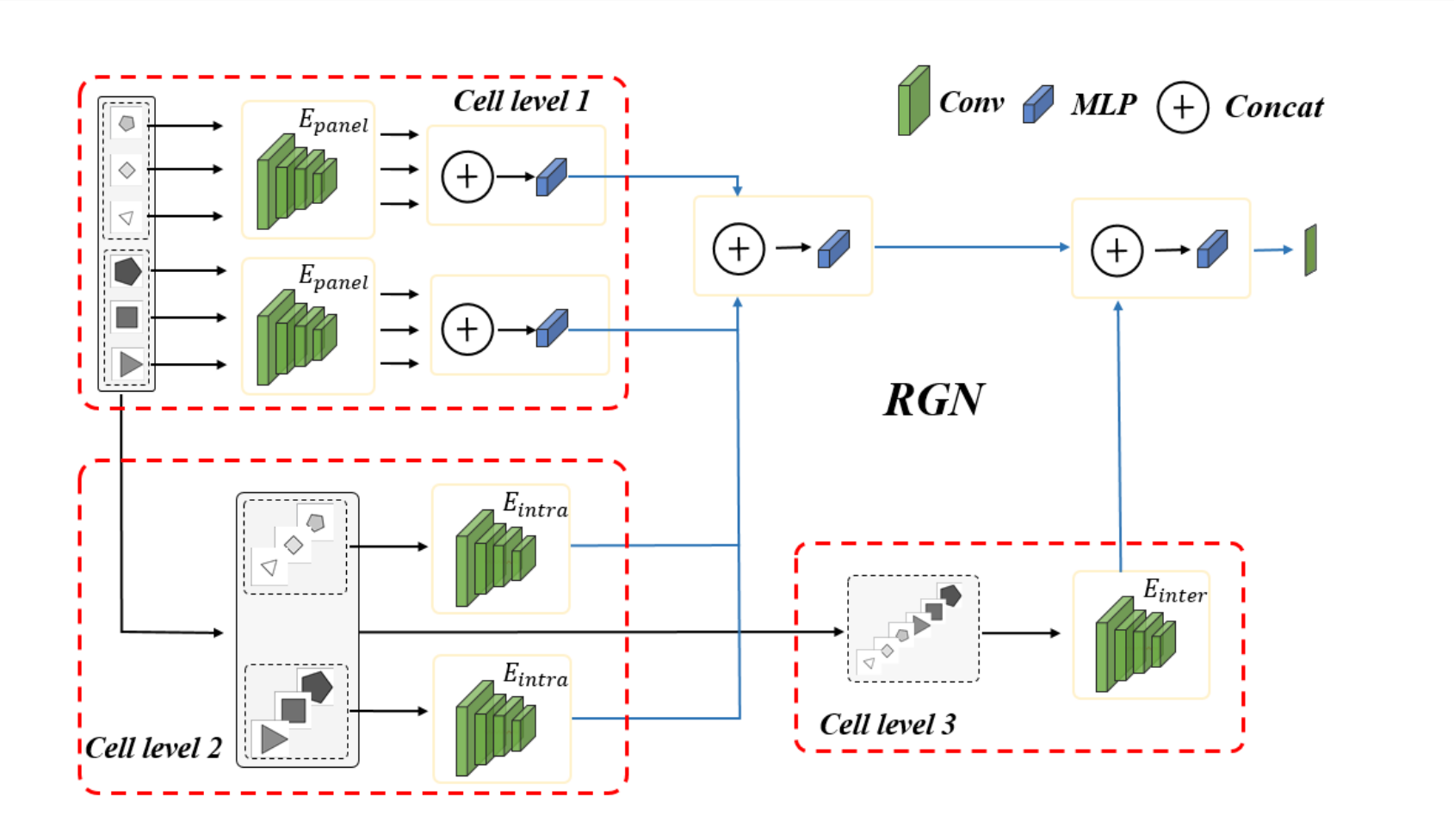
Ostatecznie $r_{ij}^{(3)}$ jest reprezentacją wierszy i oraz j. Funkcja φ łączy wszystkie wejścia w wspólną reprezentacje za pomocą sieci MLP.

Model HriNet

Trenowanie modelu polega na:

- zbudowaniu dominującej reguły: $g = r_{12}^{(3)}$, na podstawie dwóch pierwszych wierszy
- wyliczeniu reguły odpowiadającej każdej odpowiedzi $k \in \{1, 2, \ldots, 8\}$ poprzez uśrednienie: $r_k = \frac{1}{2}(r_{13}^{(3)} + r_{23}^{(3)})$
- maksymalizacja podobieństwa reguły odpowiadającej dobrej odpowiedzi do dominującej reguły, w porównaniu do reguł złych odpowiedzi, za pomocą (N+1)-tuplet loss

Model HriNet



Rysunek 17: Hierarchical Rule Induction Network.

Wyniki na zbiorze Balanced-RAVEN

Rysunek 18: Szczegółowe wyniki modeli na zbiorze Balanced-RAVEN.

Method	Acc (%)	Center (%)	2x2Grid (%)	3x3Grid (%)	L-R (%)	U-D (%)	O-IC (%)	O-IG (%)
LSTM	18.9	26.2	16.7	15.1	14.6	16.5	21.9	21.1
ResNet	40.3	44.7	29.3	27.9	51.2	47.4	46.2	35.8
ResNet+DRT	40.4	46.5	28.8	27.3	50.1	49.8	46.0	34.2
Wild ResNet	44.3	50.9	33.1	30.8	53.1	52.6	50.9	38.7
WReN	23.8	29.4	26.8	23.5	21.9	21.4	22.5	21.5
WReN (ResNet)	42.6	75.7	45.9	39.0	31.2	34.8	37.2	34.8
HriNet	63.9	80.1	53.3	46.0	72.8	74.5	71.0	49.6

Podsumowanie

Główne pomysły w pracy magisterskiej

- Sieć relacyjna działająca na wybranych trójkach obiektów zamiast na wszystkich parach
- Porównywanie pojedyńczych obiektów zamiast całych paneli
- Użycie modułu atencji do skupiania się tylko na ważnych relacjach

Wnioski z dotychczasowych prac

- Problemy analizy relacji w obrazach wymagają skupienia się na architekturze i sposobie trenowania zamiast na przygotowaniu danych
- Dużo otwartych problemów gdzie istnieją tylko baseline'y

Dziękuję za uwagę

Bibliografia i

J. Mańdziuk and A. Żychowski .
Deepiq: A human-inspired ai system for solving iq test problems.
In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2019.

David G. T. Barrett, Felix Hill, Adam Santoro, Ari S. Morcos, and Timothy P. Lillicrap.
Measuring abstract reasoning in neural networks.
CoRR, abs/1807.04225, 2018.

Bibliografia ii

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai.
Hierarchical rule induction network for abstract visual reasoning.
arXiv preprint arXiv:2002.06838, 2020.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei - Fei, C. Lawrence Zitnick, and Ross B. Girshick.
CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning.
CoRR, abs/1612.06890, 2016.

Bibliografia iii


Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap.
A simple neural network module for relational reasoning.
CoRR, abs/1706.01427, 2017.


Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merri ë nboer, and Tomas Mikolov.
Towards ai-complete question answering: Aset of prerequisite toy tasks.
arXiv preprint arXiv:1502.05698, 2015.

Bibliografia iv


Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song - Chun Zhu.
RAVEN: A dataset for relational and analogical visual reasoning.
CoRR, abs/1903.02741, 2019.