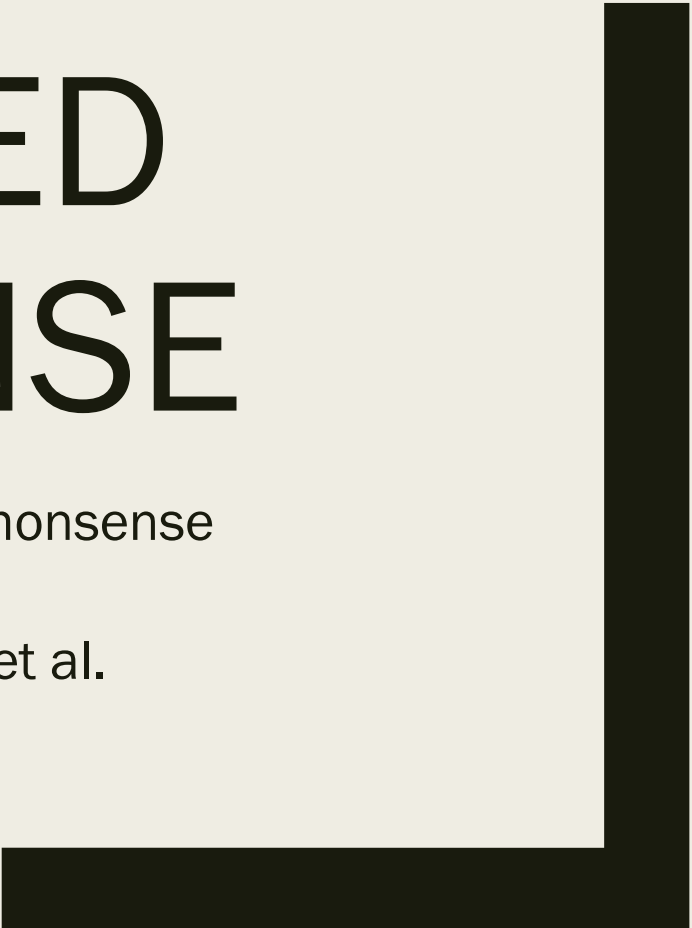


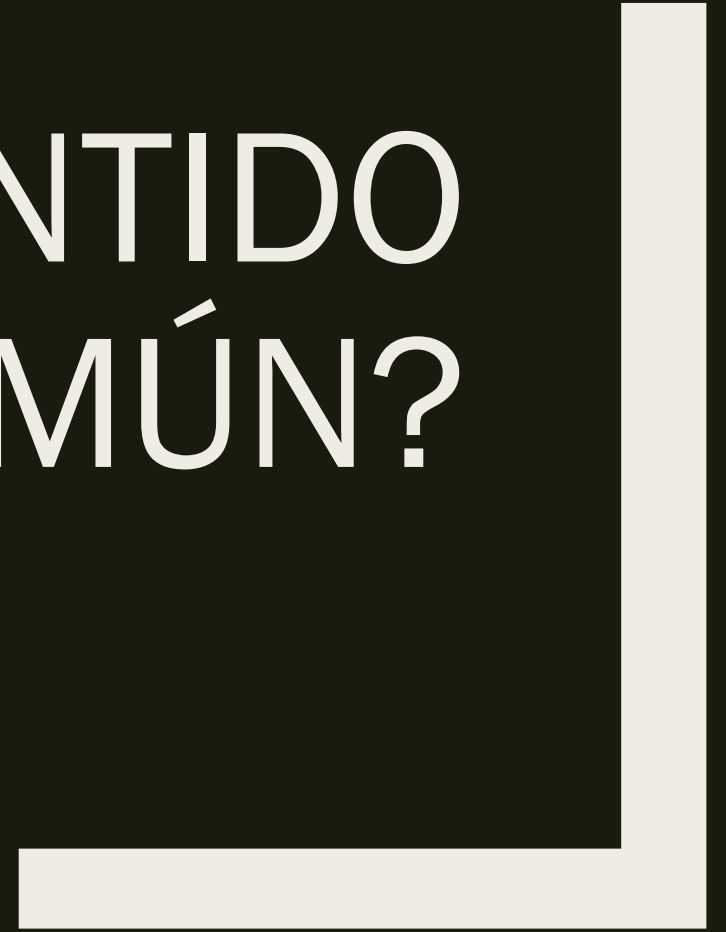
SHORTCUTTED COMMONSENSE

Data Spuriousness in Deep Learning of Commonsense
Reasoning

Autores: Ruben Branco, António Branco, et al.



¿QUÉ ES EL SENTIDO
COMÚN?



Sentido Común

- Rendimiento en Argument Reasoning Comprehension Test

Sentido Común

- Rendimiento en Argument Reasoning Comprehension Test

77%

Bert

Sentido Común

- Rendimiento en Argument Reasoning Comprehension Test

77%
Bert

80%
Humano sin
entrenamiento

Sentido Común

IS

Sentido Común

IS

DO

Sentido Común

IS

DO

ARE

Sentido Común

IS

DO

ARE

Single NOT

Shortcuts

- Generalizaciones incorrectas usando datos falsos o erróneos

Shortcuts

- Generalizaciones incorrectas usando datos falsos o erróneos
- No leen mucho, usando pequeñas frases también alcanzan alto rendimiento

Data Contamination

- Modelos pueden acceder datos interiorizados

Data Contamination

- Modelos pueden acceder datos interiorizados
- Es un problema si los sets de datos y testeo comparten texto

¿REALMENTE TIENEN
SENTIDO COMÚN?



Tasks

Argument Reasoning Comprehension

- Habilidades de razonamiento
- Requiere más que solo lógica

ARCT Example

Reason: People choose not to use Google.

Claim: Google is not a harmful monopoly.

Warrant 1: all other search engines
re-direct to Google.

Warrant 2: other search engines do not
re-direct to Google.

Correct warrant: 2

Tasks

Argument Reasoning Comprehension

- Habilidades de razonamiento
- Requiere más que solo lógica

ARCT Example

Reason: People choose not to use Google.
Claim: Google is not a harmful monopoly.

Warrant 1: all other search engines re-direct to Google.
Warrant 2: other search engines do not re-direct to Google.

Correct warrant: 2

AI2 Reasoning Challenge

- Preguntas de alternativas de ciencias básicas

ARC Example

Question: Air has no color and cannot be seen, yet it takes up space. What could be done to show it takes up space?

Answer A: observe clouds forming.
Answer B: measure the air temperature.
Answer C: blow up a beach ball or balloon.
Answer D: weigh a glass before and after it is filled with water.

Correct answer: C

Tasks

Physical Interaction QA

- Preguntas de sentido común de física

PIQA Example

Goal: What can I use to help filter water when I am camping.

Solution 1: You can use a water filtration system like a Brita pitcher.

Solution 2: Coffee filters are a cheap and effective method to filter water when outdoors.

Correct solution: 2

Tasks

Physical Interaction QA

- Preguntas de sentido común de física

PIQA Example

Goal: What can I use to help filter water when I am camping.

Solution 1: You can use a water filtration system like a Brita pitcher.
Solution 2: Coffee filters are a cheap and effective method to filter water when outdoors.

Correct solution: 2

CommonsenseQA

- Preguntas múltiples de sentido común
- Preguntas espaciales, causa y efecto

CSQA Example

Question: What is something someone driving a car needs even to begin?

Answer A: practice.
Answer B: feet.
Answer C: sight.
Answer D: keys.
Answer E: open car door.

Correct answer: C

Metodología

- Fine-tunear los modelos a estos datasets

Metodología

- Fine-tunear los modelos a estos datasets
- Probar **inputs incompletos** y comparar el rendimiento

Metodología

- Fine-tunear los modelos a estos datasets
- Probar **inputs incompletos** y comparar el rendimiento
- Realizar **ataques adversarios**, donde los tests se modifican levemente, pero manteniendo el significado

Metodología

- Fine-tune los modelos a estos datasets
- Probar **inputs incompletos** y comparar el rendimiento
- Realizar **ataques adversarios**, donde los tests se modifican levemente, pero manteniendo el significado
- Evaluar la **contaminación de datos** entre tests de pre-entrenamiento y los tasks

Metodología

- Fine-tunear los modelos a estos datasets
- Probar **inputs incompletos** y comparar el rendimiento
- Realizar **ataques adversarios**, donde los tests se modifican levemente, pero manteniendo el significado
- Evaluar la **contaminación de datos** entre tests de pre-entrenamiento y los tasks
- Probar la habilidad de generalización **cruzando tareas**

Modelos

- RoBERTa, solo encoder
- GPT-2, solo decoder
- T5, encoder-decoder
- Bart, como baseline de modelos neuro simbólicos
- COMET(BART), modelo neuro simbólico donde BART se entrena sobre una base de sentido común usando tareas generativas

RESULTADOS BASE

	ARCT	ARC	PIQA	CSQA	Params
Random	0.5	0.25	0.5	0.2	-
HUMAN	0.909	N/A	0.949	0.889	-
RoBERTa-Large	0.815 ± 0.011	0.411 ± 0.022	0.789 ± 0.006	0.733 ± 0.006	355M
GPT2-Medium	0.540 ± 0.071	0.318 ± 0.009	0.706 ± 0.005	0.551 ± 0.012	345M
T5-Large	0.743 ± 0.006	0.440 ± 0.008	0.772 ± 0.005	0.713 ± 0.007	770M
BART-Large	0.655 ± 0.154	0.382 ± 0.027	0.777 ± 0.005	0.738 ± 0.005	406M
COMET(BART)	0.790 ± 0.005	0.412 ± 0.011	0.783 ± 0.008	0.718 ± 0.008	406M

RESULTADOS BASE

	ARCT	ARC	PIQA	CSQA	Params
Random	0.5	0.25	0.5	0.2	-
HUMAN	0.909	N/A	0.949	0.889	-
RoBERTa-Large	0.815 ± 0.011	0.411 ± 0.022	0.789 ± 0.006	0.733 ± 0.006	355M
GPT2-Medium	0.540 ± 0.071	0.318 ± 0.009	0.706 ± 0.005	0.551 ± 0.012	345M
T5-Large	0.743 ± 0.006	0.440 ± 0.008	0.772 ± 0.005	0.713 ± 0.007	770M
BART-Large	0.655 ± 0.154	0.382 ± 0.027	0.777 ± 0.005	0.738 ± 0.005	406M
COMET(BART)	0.790 ± 0.005	0.412 ± 0.011	0.783 ± 0.008	0.718 ± 0.008	406M

RESULTADOS BASE

	ARCT	ARC	PIQA	CSQA	Params
Random	0.5	0.25	0.5	0.2	-
HUMAN	0.909	N/A	0.949	0.889	-
RoBERTa-Large	0.815 ± 0.011	0.411 ± 0.022	0.789 ± 0.006	0.733 ± 0.006	355M
GPT2-Medium	0.540 ± 0.071	0.318 ± 0.009	0.706 ± 0.005	0.551 ± 0.012	345M
T5-Large	0.743 ± 0.006	0.440 ± 0.008	0.772 ± 0.005	0.713 ± 0.007	770M
BART-Large	0.655 ± 0.154	0.382 ± 0.027	0.777 ± 0.005	0.738 ± 0.005	406M
COMET(BART)	0.790 ± 0.005	0.412 ± 0.011	0.783 ± 0.008	0.718 ± 0.008	406M

Inputs Parciales

- Ambos modelos tienen buen rendimiento en ARCT y PIQA

Inputs Parciales

- Ambos modelos tienen buen rendimiento en ARCT y PIQA
- Para PIQA, usaron solo respuestas como input

PIQA Example

Goal: What can I use to help filter water when I am camping.

Solution 1: You can use a water filtration system like a Brita pitcher.

Solution 2: Coffee filters are a cheap and effective method to filter water when outdoors.

Correct solution: 2

0.795 RoBERTa

0.794 COMET(BART)

Inputs Parciales

- Ambos modelos tienen buen rendimiento en ARCT y PIQA
- Para PIQA, usaron solo respuestas como input

Solution 1: You can use a water filtration system like a brita pitcher.
Solution 2: Coffee filters are a cheap and effective method to filter water when outdoors.

0.735 RoBERTa

0.724 COMET(BART)

Inputs Parciales

- Ambos modelos tienen buen rendimiento en ARCT y PIQA
- Para PIQA, usaron solo respuestas como input
- Otros dataset mostraron problemas similares, excepto CSQA

Ataques Adversarios

Question: Ira had to make up a lab investigation after school. He obtained the materials, chemicals, equipment, and protective gear from his teacher. Quickly, but cautiously, he conducted the steps in the written experiment procedure. To save time, he decided to record his observations and results later. Which will most likely be negatively affected by his decision?

Before

- A: the ability to follow directions
- B: the ability to write a valid report
- C: the ability to follow the safety guidelines
- D: the ability to come up with a conclusion

After

- A: the capacity to follow directions
- B: the ability to write a valid report
- C: the ability to follow the safety guidelines
- D: the ability to come up with a conclusion

Correct choice: B

Model's choice: B ✓

Model's choice after perturbation: A ✗

Ataques Adversarios

Question: Ira had to make up a lab investigation after school. He obtained the materials, chemicals, equipment, and protective gear from his teacher. Quickly, but cautiously, he conducted the steps in the written experiment procedure. To save time, he decided to record his observations and results later. Which will most likely be negatively affected by his decision?

Before

- A: the **ability** to follow directions
- B: the ability to write a valid report
- C: the ability to follow the safety guidelines
- D: the ability to come up with a conclusion

After

- A: the **capacity** to follow directions
- B: the ability to write a valid report
- C: the ability to follow the safety guidelines
- D: the ability to come up with a conclusion

Correct choice: B

Model's choice: B ✓

Model's choice after **perturbation**: A ✗

Caída 27%-64% RoBERTa
Caída 31%-75% COMET(BART)

Contaminación de Datos

- Los sets no se encuentran muy contaminados

Contaminación de Datos

- Los sets no se encuentran muy contaminados
- PIQA tiene la mayor, con 13%

Contaminación de Datos

- Los sets no se encuentran muy contaminados
- PIQA tiene la mayor, con 13%
- Estudio más profundo, aparte de ARC, contaminación no da ventaja

Tareas Cruzadas

- Se busca el rendimiento zero-shot en otras tareas

Tareas Cruzadas

- Se busca el rendimiento zero-shot en otras tareas

	ARCT	ARC	PIQA	CSQA
ARCT	<i>0.831</i>	0.310	0.571	0.293
ARC	0.589	<i>0.435</i>	0.627	0.343
PIQA	0.597	0.230	<i>0.795</i>	0.552
CSQA	0.627	0.384	0.687	<i>0.738</i>
Random	0.5	0.25	0.5	0.2

Conclusión

- Los modelos tienen un rendimiento superior al esperado cuando parte relevante del input falta

Conclusión

- Los modelos tienen un rendimiento superior al esperado cuando parte relevante del input falta
- Ataques adversarios demuestran que baja el rendimiento al cambiar el texto del input, pero no la pregunta

Conclusión

- Los modelos tienen un rendimiento superior al esperado cuando parte relevante del input falta
- Ataques adversarios demuestran que baja el rendimiento al cambiar el texto del input, pero no la pregunta
- La falta de contaminación cruzada revela que los modelos no están interiorizando información

Conclusión

- Los modelos tienen un rendimiento superior al esperado cuando parte relevante del input falta
- Ataques adversarios demuestran que baja el rendimiento al cambiar el texto del input, pero no la pregunta
- La falta de contaminación cruzada revela que los modelos no están interiorizando información
- A partir de esto, se concluye que existe algún tipo de relación entre los datos que no es aparente

Crítica

- Muy interesante

Crítica

- Muy interesante
- Muy claro

Crítica

- Muy interesante
- Muy claro
- Ver más ejemplos