From Strings to Things: Knowledge-enabled VQA Model that can Read and Reason (Supplementary Material)

Ajeet Kumar Singh[1] Anand Mishra[2*] Shashank Shekhar[3] Anirban Chakraborty[3]

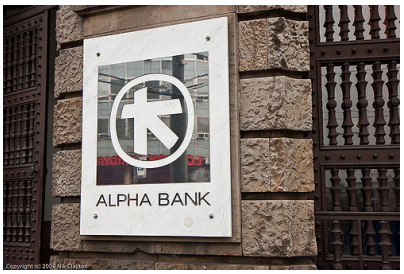
[1]TCS Research, Pune, India [2]IIT Jodhpur, India [3]Indian Institute of Science, Bangalore, India

Contents

*Anand Mishra was associated with the Indian Institute of Science, Bangalore when this work was carried out.

1. A selection of images and QA pairs from text-KVQA

1.1. Dataset sample - Page 1



(a)
Q: What is this building?
A: **Bank**

(b)
Q: Which bank is this?
A: **Bank of America**

(c)
Q: Which restaurant is this?
A: **Baskin-Robbins**

(d)
Q: Which company owns this brand?
A: **Walmart**

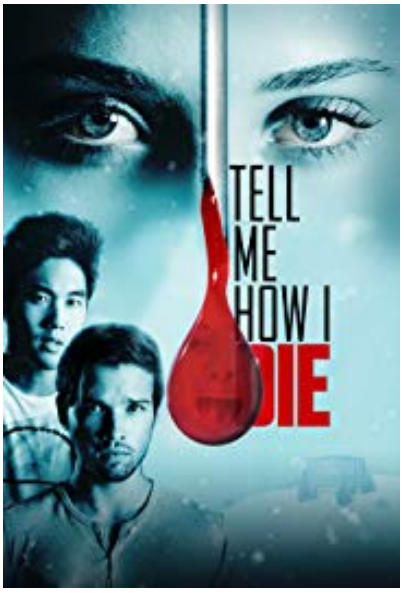
(e)
Q: Is this a car showroom?
A: **No**

(f)
Q: Is this an animation movie?
A: **Yes**

Figure 1. A selection of images and question-ground truth answer pairs from our newly introduced dataset viz. text-KVQA. This dataset will be made publicly available for future research.

1.2. Dataset sample - Page 2

(a)
Q: Can I get coffee here?
A: **Yes**



(b)
Q: Who is the director of this movie?
A: **D. J. Viola**

(c)
Q: Is this a German brand?
A: **No**

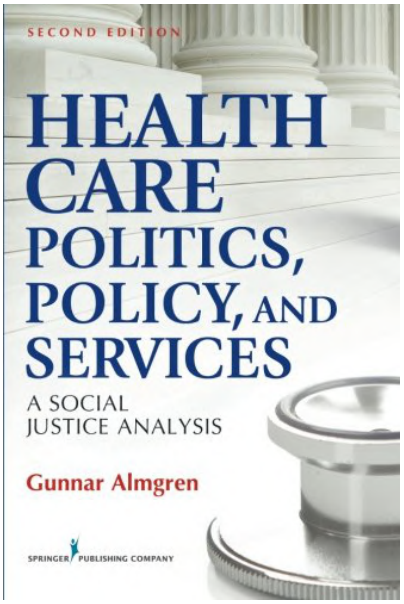
(d)
Q: Can I fill fuel in my car here?
A: **No**

(e)
Q: Is this a Portuguese brand?
A: **Yes**

(f)
Q: Which electronic store is this?
A: **Panasonic**

Figure 2. A selection of images and question-ground truth answer pairs from our newly introduced dataset viz. text-KVQA. This dataset will be made publicly available for future research.

1.3. Dataset sample - Page 3



(a)
Q: What is the genre of this book?
A: **Medical Books**

(b)
Q: Does it sell Pizza?
A: **Yes**

(c)
Q: Which restaurant is this?
A: **Cafe Coffee Day**

(d)
Q: What is this?
A: **Book store**

(e)
Q: What does this store sell?
A: **Watches**

(f)
Q: Is this a Dutch brand?
A: **Yes**

Figure 3. A selection of images and question-ground truth answer pairs from our newly introduced dataset viz. text-KVQA. This dataset will be made publicly available for future research.

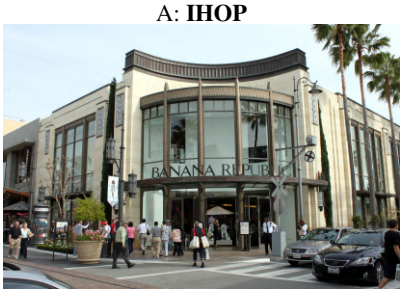
1.4. Dataset sample - Page 4

(a)
Q: What is this?
A: **Restaurant**

(b)
Q: Which restaurant is this?
A: **IHOP**

(c)
Q: Can I get medicine here?
A: **Yes**

(d)
Q: Is this a car showroom?
A: **No**



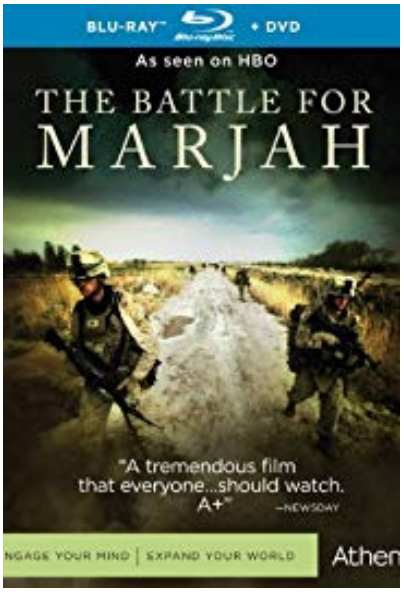
(e)
Q: What does this store sell?
A: **Clothing and Jewelry**

(f)
Q: Which mobile store is this?
A: **Vodafone**

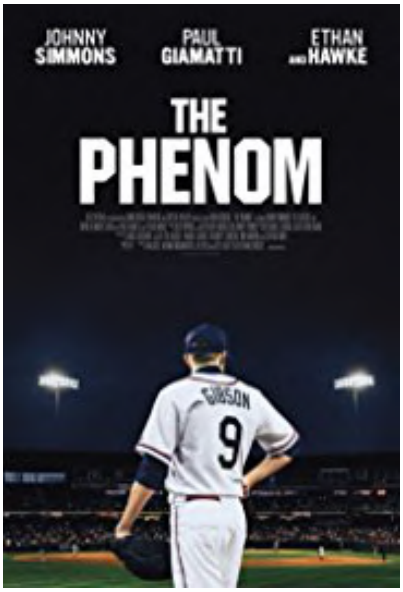
Figure 4. A selection of images and question-ground truth answer pairs from our newly introduced dataset viz. text-KVQA. This dataset will be made publicly available for future research.

1.5. Dataset sample - Page 5

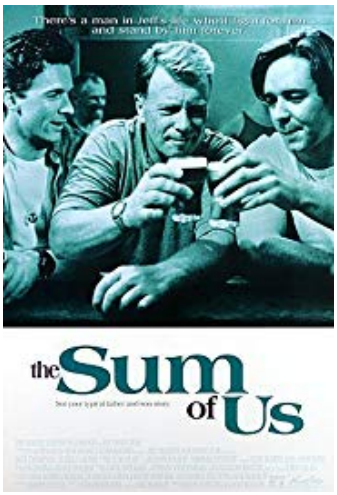
(a)
Q: Which retail store is this?
A: **Asda**



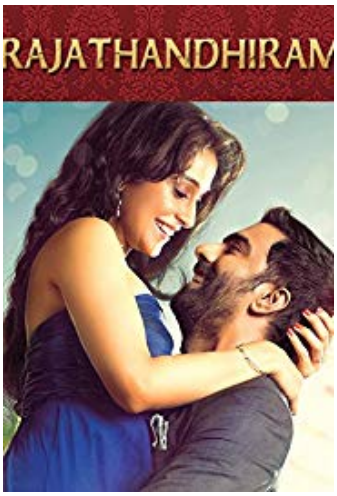
(b)
Q: What is the genre of this movie?
A: **Documentary**



(c)
Q: Who was the director of this movie?
A: **Noah Buschel**



(d)
Q: When was this movie released?
A: **1994**



(e)
Q: What is the language of this movie?
A: **Tamil**

(f)
Q: What is the title of this movie?
A: **Aftershock**

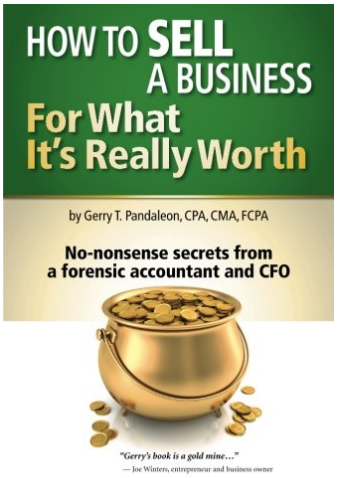
Figure 5. A selection of images and question-ground truth answer pairs from our newly introduced dataset viz. text-KVQA. This dataset will be made publicly available for future research.

1.6. Dataset sample - Page 6

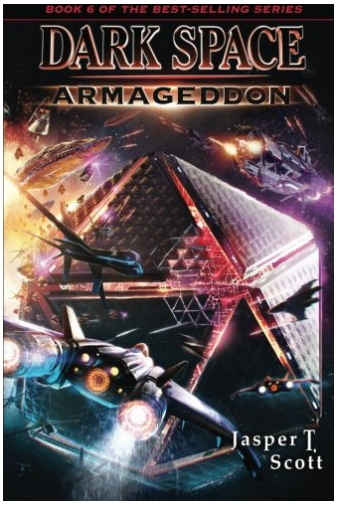
(a)
Q: Is this an animation movie?
A: **Yes**

(b)
Q: In which language this book is written?
A: **French**

(c)
Q: Who published this book?
A: **Broadman & Holman Publishers**



(d)
Q: Is this book related to finance?
A: **Yes**



(e)
Q: Who wrote this book?
A: **Jasper T. Scott**

(f)
Q: Is this an American brand?
A: **Yes**

Figure 6. A selection of images and question-ground truth answer pairs from our newly introduced dataset viz. text-KVQA. This dataset will be made publicly available for future research.

2. Detail analysis of text-KVQA

2.1. Visual contents - Page 1

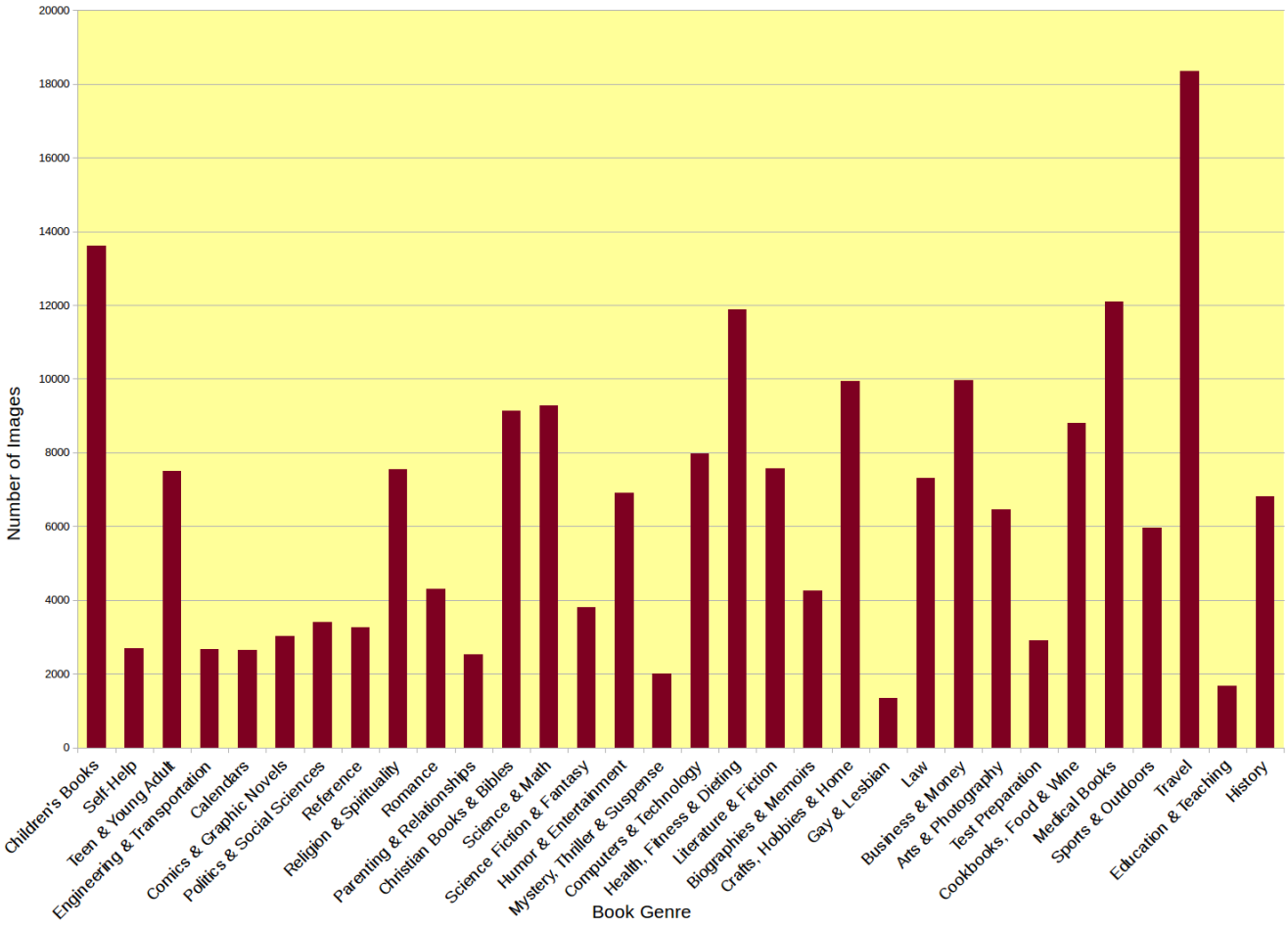


Figure 7. Visual contents in text-KVQA (book). It contains more than 200K book covers from 32 book genres in all.

2.2. Visual Contents - Page 2

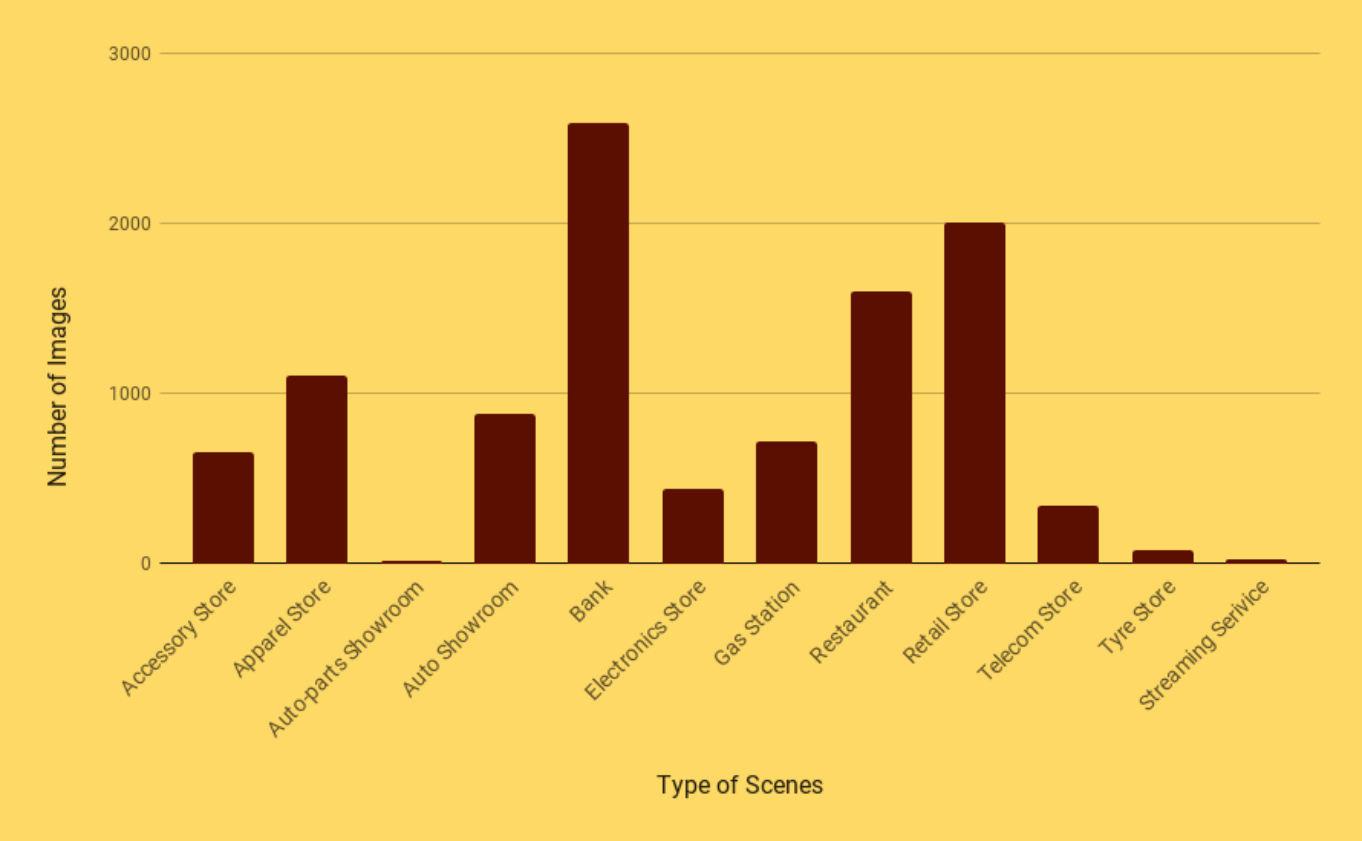


Figure 8. Visual contents in text-KVQA (scene). It contains diverse set of scenes ranging from banks, restaurants, gas stations, electronic stores, apparel stores, etc.

2.3. Visual contents - Page 3

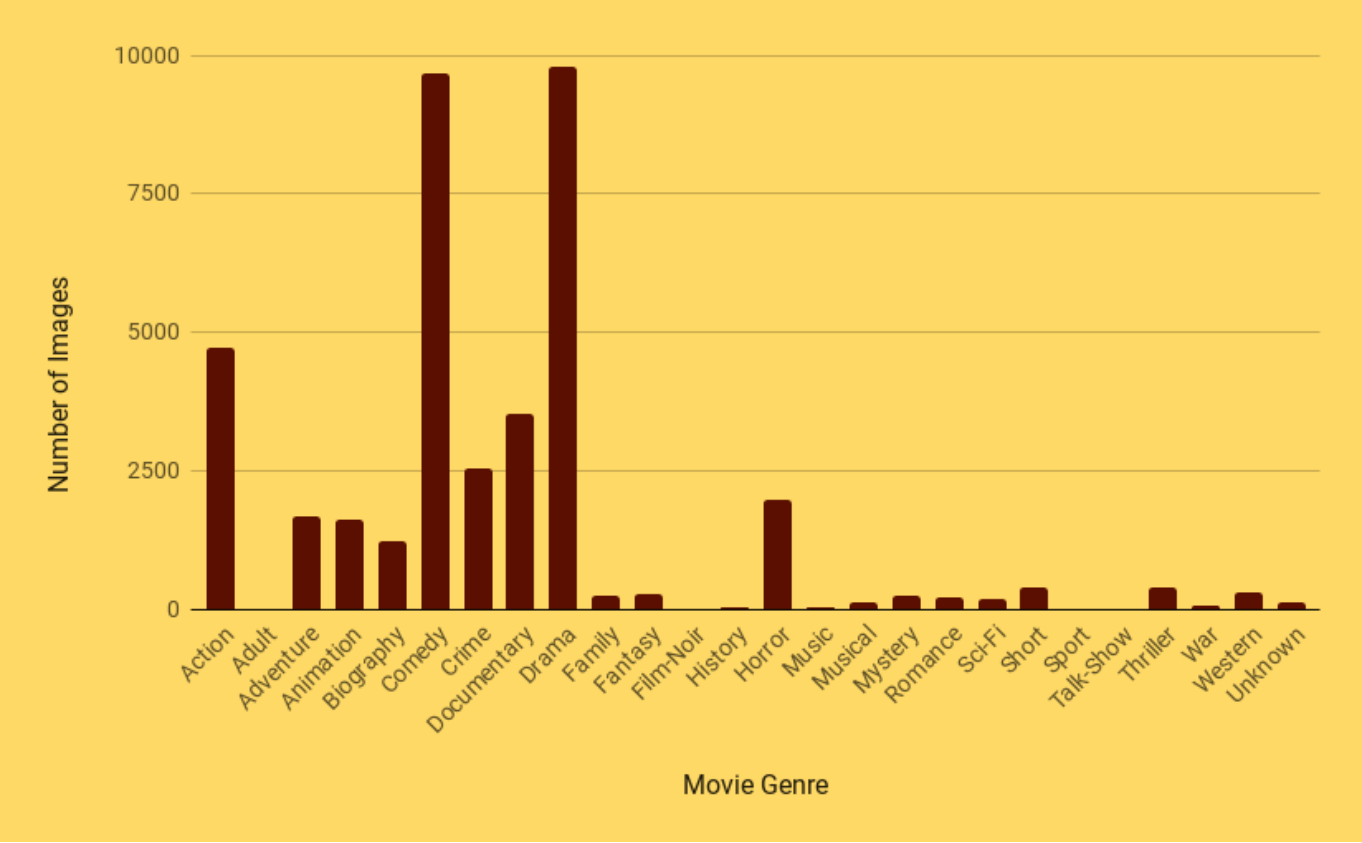


Figure 9. Visual contents in text-KVQA (movie). It contains more than 39K movies posters from 32 movie genres in all.

2.4. Answers as world cloud in text-KVQA (book)

Figure 10. Word cloud of top-100 words in text-KVQA (book). It indicates towards the diversity in answer space and balanced nature of questions.

2.5. Answers as world cloud text-KVQA (scene)

Figure 11. Word cloud of top-100 words in text-KVQA (scene). It indicates towards the diversity in answer space and balanced nature of questions.

2.6. Answer distribution in text-KVQA

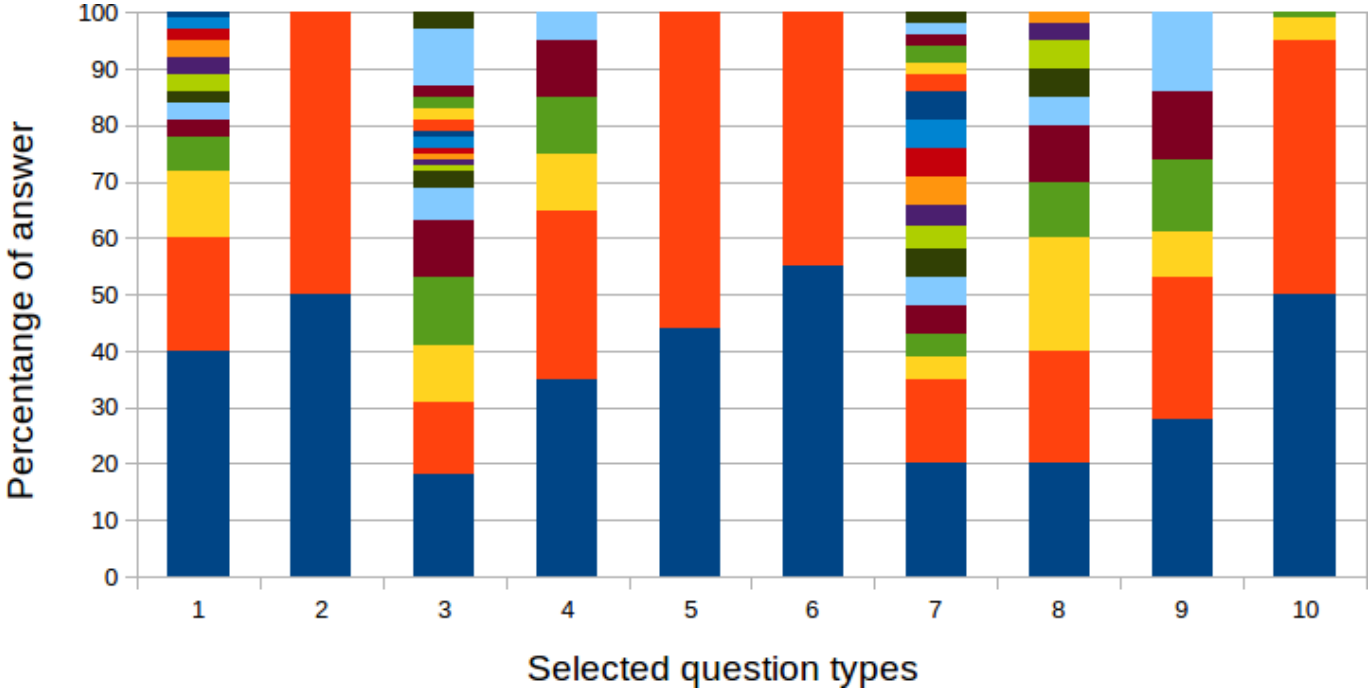


Figure 12. Answer distribution in text-KVQA for a few selected question types. 1: What is this, 2: Can I get, 3: Which restaurant, 4: What is the language, 5: Does it sell, 6: Which year calendar, 7: When was, 8: Which clothing store, 9: Which gas station, 10: Which auto-part showroom. Different color indicates different answer, for example, for question type 2 (Can I get) blue and red colors indicate answer *Yes* and *No* respectively. Many other questions, for example inquiring movie names, book titles, author names, genre types are naturally balanced due to large answer space.

3. Results of our VQA model

3.1. Results - Page 1

(a) **Detected words:** {GAP}
Word Proposals: {GAP, GALP}
Visual Content Proposal: {Clothing store, department store, gift shop}
Q: What is this store?
A: Clothing store
Supporting fact: GAP is a clothing store.
Observation: GALP is a petroleum brand, visual contents helps here to recover from lower precision in word proposals.

(b)
Detected words: {Baja, c, d}
Word Proposals: {Bata, c, d}
Visual Content Proposal: {Clothing store, Shoe shop, Gift shop}
Q: Which shoe shop is this?
A: Bata
Supporting fact: Bata is a shoe brand.
Observation: recovers from wrong recognition: Baja.

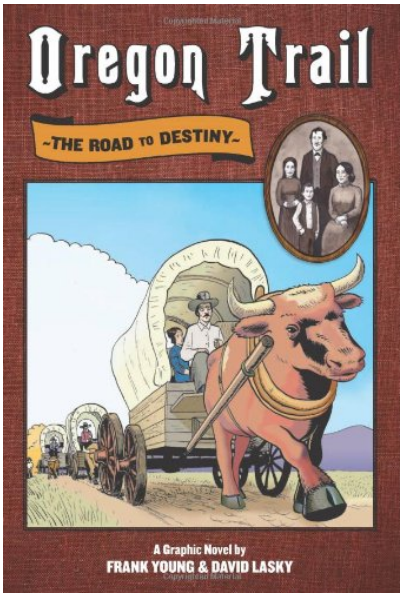
(c)
Detected words: {Arai, Arai, Arai, 11}
Word Proposals: {Aral, 11}
Visual Content Proposal: {Fastfood restaurant, Gas station, Industrial area}
Q: Is this a German brand?
A: Yes
Supporting fact: Aral is brand of Germany.
Observation: Top-1 place recognition goes wrong here, but word proposal helps.

Figure 13. Results of our VQA model. Observation is noted with every image, question, answer tuple. Orange, blue and red colors show anchor entity, correct and incorrect answer respectively. **[Best viewed in color]**.

3.2. Results - Page 2

(a)
Detected words: {Lears}
Word Proposals: {Sears}
Visual Content Proposal: {Clothing store, Fastfood restaurant, Jewelery shop}
Q: Does it sell cloths?
A: Yes
Supporting fact: Sears is a clothing brand. **Observation:** Both word proposal and visual content mislead.

(b)
Detected words: {Luminosity, Shill}
Word Proposals: {Luminosity, Shell}
Visual Content Proposal: {Gas station, Fire station, General store}
Q: Can I fill fuel in my car here?
A: Yes
Supporting fact: Shell sells gas.
Observation: Word and visual content both misleads.



(c)
Detected words: {Oregon, Trail, The, Road, to, Destiny, Frank, Young, David, Laski}
Word Proposals: {Oregon Trail: The Road to Destiny, Young, David, Laski}
Visual Content Proposal: {Children's Books, Arts and Photography, Travel}
Q: What is the title of this book?
A: Oregon Trail: The Road to Destiny
Supporting fact: Oregon Trail: The Road to Destiny is a Children's Books.
Observation: Works even for long answers.

Figure 14. Results of our VQA model. Observation is noted with every image, question, answer tuple. Orange, blue and red colors show anchor entity, correct and incorrect answer respectively. **[Best viewed in color]**.

4. Comparison of text-KVQA with other VQA datasets

Dataset	Number of images	Number of QA pairs	Knowledge-enabled
text-KVQA (This Work)	**257,380**	**1,322,272**	**Yes**
ST-VQA [1])	23,038	31,791	No
OCRVQA-200K [2]	207,572	1,002,146	No
text-VQA [3]	28,408	45,336	No

Table 1. text-KVQA as compared to related datasets which identifies the need for reading text for VQA task. Our dataset is not only significantly larger than these datasets, but also only dataset which identifies the need for background knowledge in answering questions.

5. Details for GGNN Training

We set the hidden state dimension to 110 and number of time steps to 5. The output network is a 2-layer fully connected network. In this, first layer activation is set to *Sigmoid* and second layer to *tanh*. The initial learning rate, momentum, batch size and maximum number of epochs is set to 0.1, 0.9, 16 and 100 respectively. Learning rate is decreased by a factor of 0.1 at every 10 epochs.

References

[1] Ali Furkan Biten, Ruben Tito, Andrés Mafla, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. *CoRR*, abs/1905.13648, 2019.

[2] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*, 2019.

[3] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019.