Extract triples and organize (from multimodal input)

Train to fill missing words, etc.

Part-I

Part-II

Part-III

Task

Evaluate KG

symbolic query, matching reqd

knowledge triple(s) retrieved

decode query relevant knowledge

embed symbolic query, auto matching

symbolic representation

symbolic triple(s) retrieved

decode query relevant knowledge

symbolic query, matching reqd

knowledge triple(s) retrieved

decode query relevant knowledge

embed symbolic query, auto matching
Multimodal KGs: NEIL KB

Scene-object relationships mined

- Helicopter is found in Airfield
- Zebra is found in Savannah
- Ferris wheel is found in Amusement park
- Throne is found in Throne room
- Leaning tower is found in Pisa
- Opera house is found in Sydney
- Bus is found in Bus depot outdoor
- Camry is found in Pub outdoor

Object-object relationships mined

- Van is a kind of looks similar to Ambulance
- Eye is a part of Baby
- Duck is a kind of looks similar to Goose
- Gypsy moth is a kind of looks similar to Butterfly
- Monitor is a kind of looks similar to Desktop computer
- Sparrow is a kind of looks similar to bird
- Basketball net is a part of Backboard

visual knowledge complements typical textual KG e.g. “monitor is expensive”

NEIL: Extracting Visual Knowledge from Web Data, Chen et. al, ICCV 2013
NEIL KB: Approach

(0) Seed Images

Desktop Computer
Monitor
Keyboard
Television

(1) Visual Cluster Discovery (Section 3.1)

(2) Train Detectors

(3) Relationship Discovery (Section 3.2)

(4) Add New Instances (Section 3.3)

(5) Retrain Detectors

Learned facts:
- Monitor is a part of Desktop Computer
- Keyboard is a part of Desktop Computer
- Television looks similar to Monitor

NEIL: Extracting Visual Knowledge from Web Data, Chen et. al, ICCV 2013
GB-NET: from scene graphs to CSK graphs

Scene graphs are image dependent

Commonsense graphs are image independent

Bridging Knowledge Graphs to Generate Scene Graphs, Zareian et. al, ECCV 2020
Situation with grounding data: SWiG

action specific tuples (frames)
### Agenda

**Part-I**
- Extract triples and organize
- Visual commonsense knowledge

**Part-II**
- Symbolic representation
- Neural helps symbolic
- Symbolic helps neural

**Part-III**
- Evaluate KG
- Knowledge triple(s) retrieved
- Symbolic query, matching reqd

**Task**
- Decode query relevant knowledge
- Embed symbolic query, auto matching

---

**Rich complementary knowledge**

**Visual vs textual knowledge:**
- Visual KG captures unmentioned knowledge.
- Might also suffer from reporting bias.

**Future research directions:**
- Extract (interaction) knowledge from videos.
- More never-ending approaches like NEIL.
From Knowledge base construction to Deep learning

Extract triples and organize

Train to fill missing word etc.

Task
- symbolic query, matching reqd
- knowledge triple(s) retrieved
- decode query relevant knowledge
- embed symbolic query, auto matching

symbolic representation
- neural representation of encoded knowledge in the hidden layers
Part-I

Extract triples and organize

Part-II

symbolic representation

Neural helps symbolic

symbolic helps neural

neural representation of encoded knowledge in the hidden layers

Part-III

Evaluate KG

Task

symbolic query, matching reqd

knowledge triple(s) retrieved

decode query relevant knowledge

embed symbolic query, auto matching

Train models

Agenda
5 min tour de Neural Language models

**Task: 😛😕**
Conferences make you want to attend them

**Transformer architecture**

**Task: typing assist**

**Bidirectional Encoder**

**Autoregressive Decoder**

**Text corpus**

**Training**

**Inference**

softmax

feed forward

finetune first

infer later

feed forward
Tour de Transformers

Transform to a really good hidden representation

different layers might capture different low/high level aspects such as texture, color, shape, size or emotion, gender

Credit: All the nice Transformer illustrations taken from http://jalammar.github.io/illustrated-transformer/
tour de Encoders in transformer
transform to a really good hidden representation

Credit: All the nice Transformer illustrations taken from http://jalammar.github.io/illustrated-transformer/
tour de Encoders in transformer
transform to a really good hidden representation

Credit: All the nice Transformer illustrations taken from http://jalammar.github.io/illustrated-transformer/
tour de Encoders in transformer
transform to a really good hidden representation
(de) tour de models

**current state of the art models**: T5 (encoder + decoder architecture) and GPT3

### Transformer architecture

- **Bidirectional Encoder**
  - Input: text corpus
  - Output: training

- **Autoregressive Decoder**
  - Input: I like to attend tutorials
  - Output: P(x | I like)

#### Tasks
- **Task**: 😁😕
  - Conferences make you want to attend them

- **Task**: typing assist
  - Feed forward softmax
  - Infer later

---

**BERT**

- Input: I like to attend tutorials
- Output: Transformer architecture

---

**GPT**

- Input: I like to attend tutorials
- Output: P(x | I like)
Extract triples and organize

Train models

Part-I

Part-II

symbolic representation

neural representation of encoded knowledge in the hidden layers

Part-III

Evaluate KG

Part-III Agenda

1. Task

symbolic query, matching reqd

knowledge triple(s) retrieved

decode query relevant knowledge

embed symbolic query, auto matching

Neural helps symbolic

Neural helps symbolic

symbolic helps neural

Feed Forward Neural Network

Train models

Part-I

Part-II

Part-III
untuned model\textsuperscript{3} is not great

**Tokens:** [CLS] everyone knows that a bear has [MASK]. [SEP]

**All Results:**
1: teeth - 34.521595%
2: fangs - 15.836702%
3: wings - 5.113015%
4: horns - 4.042341%
5: claws - 3.797797%
6: eyes - 3.060219%
7: legs - 2.741149%
8: fur - 1.653655%
9: ears - 1.173016%

tuned model\textsuperscript{4} is much better (like with any neural LM)

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response</th>
<th>PLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a</td>
<td>fur</td>
<td>27</td>
<td>teeth</td>
<td>.36</td>
</tr>
<tr>
<td>bear has ___</td>
<td>claws</td>
<td>15</td>
<td>claws</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>teeth</td>
<td>11</td>
<td>eyes</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>cubs</td>
<td>7</td>
<td>ears</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>paws</td>
<td>7</td>
<td>horns</td>
<td>.02</td>
</tr>
</tbody>
</table>

low correlation with human elicit properties but are coherent.

can also distinguish based on properties: “X has fur” vs “X has fur and is big”

\textsuperscript{3} Bar ilan demo., as of 2021: link

\textsuperscript{4} Weir et al., 2020

\textsuperscript{5} Forbes et al., 2019
"neural language representations still only learn associations that are explicitly written down"\(^5\), even after being explicitly trained on a knowledge graph of objects and affordances.

"Perceptual or visual concepts such as smooth, can’t be learned from text alone"\(^4\), based on properties: “X has fur” vs “X has fur and is big”
2 of 4: multi-relational & visual knowledge in neural LMs

autoregressive model\(^2\)
(GPT fine-tuned on ConceptNet)

[1] (again, untuned is bad)
AllenNLP demo GPT2, as of 2021: link

[2] COMeT demo., Bosselut et. al, as of 2021: link
“Do not handle mutual exclusivity well and suffer from frequency bias (in general the outputs may be incoherent or inconsistent)”

“Perceptual or visual concepts still hard to learn”
Task: Generate events before, after and intents at present given an image, and a description of the event in the image, and a plausible scene/location. Uses visual and language transformer.

Event: [Person2] is holding onto a bronze statue while waves of water crash around him.
Place: Inside a sinking ship

Before Person2 needed to ...
- Swim towards the statue.
- Sense his own death.
- Notice water washing in.

Because Person2 wanted to ...
- Wait for help to arrive.
- Hold on for his life.
- Save himself from drowning.

After Person2 will most likely ...
- Gasp for air.
- Scream for help.
- Be washed away.

[6] Park et al., ECCV 2020
https://visualcomet.xyz
Task: Generate events before, after and intents at present given an image, and a description of the event in the image, and a plausible scene/location.

Great start, future models could make fewer assumptions.

Output can still be inconsistent and incoherent.

[6] Park et al., ECCV 2020
https://visualcomet.xyz
Similar ideas have been applied to correct a KG based on neural LM perplexity.

However, LMs can generate fictitious facts (distributionally similar but factually wrong).

Table 1: Example of generating candidate sentences. Several enumerated sentences for the triple (musician, CapableOf, play musical instrument). The sentence with the highest log-likelihood according to a pretrained language model is selected.
Entity linkage: linking multiple taxonomies online is a massive, unsolved task.
4 of 4: neural LMs to fuse use multiple CKGs

- Entity linkage: linking multiple taxonomies online is a massive, unsolved task.
- Attention: need to first retrieve relevant subgraph.
- Multi-task learning: scalable, and embeds knowledge (e.g., UNICORN)

---

Entire KG (verbalized triples) is learned to complete as a task. So model trained on QA as well as KG prediction task.

No KG, model only trained on QA task.

---

<table>
<thead>
<tr>
<th>Knowledge Graph</th>
<th>SocialIQa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic</td>
<td>75.0</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>74.3</td>
</tr>
<tr>
<td>Both</td>
<td>74.8</td>
</tr>
<tr>
<td>Single task</td>
<td>73.8</td>
</tr>
</tbody>
</table>

UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark, Lourie et al., AAAI 2020
Pros/cons of using neural over symbolic KGs

Pros:

1. Real tasks/queries representation space might be different, and it is difficult to align with the representation space/or query the KG.
2. Typically, KGs do not come with context. This makes the KG lookup even more difficult. For example, things break when they fall but soft things do not.
Pros/cons of using neural over symbolic KGs

Cons:

1. Symbolic KGs are more interpretable and easily debuggable, but neural models are hard to probe.
2. Promising direction of multi-task learning for using multiple KGs, but more work is needed.
3. LMs can generate fictitious facts-- this requires more work. e.g., grounding the knowledge to an established source such as Wikipedia.
4. More work is required (BOTH in symbolic and neural) to acquire perceptually grounded/ unmentioned knowledge, e.g, visual COMeT with fewer assumptions in the input -- and we need to make the output more consistent.
Part-I: Extract triples and organize

Part-II: Train models

Part-III: Evaluate KG

Task: symbolic query, matching reqd
- knowledge triple(s) retrieved
- decode query relevant knowledge
- embed symbolic query, auto matching

Part-I: neural representation of encoded knowledge in the hidden layers
- Train models

Part-II: symbolic representation
- Neural helps symbolic

Part-I: Agenda

Part-II: ✓

Part-III: ✓
Can CSK help neural models

Robustness\textsuperscript{[d1]}

Generating adversarial examples guided by commonsense knowledge\textsuperscript{[d2]}

Explainability\textsuperscript{[d3]}

Using attention map generated by a QA model (top right) to identify relevant components of a scene graph\textsuperscript{[d4]}

\textsuperscript{[d1]}: Cycle-Consistency for Robust Visual QA, Shah et. al 2019
\textsuperscript{[d2]}: AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples, Kang et. al 2018
\textsuperscript{[d3]}: Generating Natural Language Explanations for Visual QA Using Scene Graphs and Visual Attention, Ghosh et al., 2018
\textsuperscript{[d4]}: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, Krishna et. al, 2016
Can CSK help neural models

Limited training data

Inject commonsense knowledge to compensate for limited training data

Difficult to find training data for all types of scenarios, esp. rarely mentioned rules and facts

- Are shiny surfaces typically hard?
- What's bigger the moon or a wolf?
- If I put my socks in the drawer, will they still be there tomorrow?
Injecting commonsense knowledge into DL models

Incorporating commonsense knowledge into deep learning (DL) models can improve their performance by providing additional information. This can be achieved through various methods, such as scoring, attention, multitask learning, and augmented input. These methods can be used to inject commonsense knowledge into DL models in the form of triples, rules, or generated knowledge graphs (KGs).

- **Free-form triples**: Informal statements that capture commonsense knowledge.
- **Canonical triples**: Structured statements that are formalized and standardized.
- **Verbalized triples**: Statements that are expressed in natural language.
- **Preconstructed KG**: A knowledge graph that is already available and can be used directly.
- **On-the-fly generated KG**: A knowledge graph that is generated during the inference process.

These methods can be applied to different tasks, such as natural language processing (NLP) and computer vision (CV), to enhance their capabilities and accuracy.
During beam search decoding to find globally consistent results, probability mass moves away from implausible states. Model has seen insufficient data to learn these correlations, so use commonsense to steer away from unrealistic states.

- Things cannot move if they don’t exist yet
- Ball will be destroyed at step 2 (less likely)

3 rules from SUMO ontology

KB
Be Consistent! Improving Procedural Text Comprehension using Label Consistency. Du et al NAACL 2019

Adds consistency loss across paragraphs (derivable from a CKG of paragraphs) while training an end2end model.
There is a recent thrust towards **unstructured entity specific sentence KGs**. It resolves the IR issues, and text can represent more complex commonsense knowledge.

1. Example generics about “tree” in GenericsKB
   - Trees are perennial plants that have long woody trunks.
   - Trees are woody plants which continue growing until they die.
   - Most trees add one new ring for each year of growth.
   - Trees produce oxygen by absorbing carbon dioxide from the air.
   - Trees are large, generally single-stemmed, woody plants.
   - Trees live in cavities or hollows.
   - Trees grow using photosynthesis, absorbing carbon dioxide and releasing oxygen.

![Diagram](image)

**Figure 1**: We study the task of open-ended commonsense reasoning (OpenCSR), where answer candidates are not provided (as in a multiple-choice setting). Given a question, a reasoner uses multi-hop reasoning over a knowledge corpus of facts, and outputs a ranked list of concepts as answers.
On the fly KG “generation” is another recent direction. When the KG is augmented to the input, QA performance boosts.

RQ1. St-Graph Generation:

Context:
Sunlight strikes chlorophyll. Sunlight trapped ...

Situation (st):
more sunlight

QA pairs:
Q2: What hurts st imminently? A2: cloudy skies
Q3: What’s helped eventually? A3: taller plants
On the fly KG “generation” is another recent direction. When the KG is augmented to the input, QA performance boosts.
Part-I

Extract triples and organize

✅ Train models

Part-II

symbolic representation

Neural helps symbolic

neural representation of encoded knowledge in the hidden layers

Part-III

Evaluate KG

symbolic query, matching reqd

knowledge triple(s) retrieved

decode query relevant knowledge

embed symbolic query, auto matching

Task

Part-I

Extract triples and organize

✅ Train models

Part-II

symbolic representation

Neural helps symbolic

neural representation of encoded knowledge in the hidden layers

Part-III

Evaluate KG

symbolic query, matching reqd

knowledge triple(s) retrieved

decode query relevant knowledge

embed symbolic query, auto matching

Task
Commonsense for Interactive learning (LeapOfThought)
inference time (current models make mistakes that can be corrected)

Ask the AI a yes/no question

Does a whale have bellybutton?

AI answer: no

AI answer: yes

Add rules to teach the AI if it answered incorrectly!

Whale is a mammal.
Commonsense for Interactive learning (LeapOfThought)
inference time (current models make mistakes that can be corrected)

- + Clearly shows that models will lack CSK and will benefit from having it.
- Model throws away the valuable user feedback after using locally.
- (risk) Model may learn false or fake information if the user tricks it.
Generating required commonsense on the fly by querying LM

Question Generation:
Because Brett found an internship while in college but Ian was unable to, what is the purpose of ___ found a job less quickly after graduation.

What is the purpose of ___ is

Answer Generation:
Because Brett found an internship while in college but Ian was unable to, what is the purpose of ___ found a job less quickly after graduation.

What is the purpose of the internship?
The purpose of the internship is

The purpose of the internship is to help people find jobs.

Unsupervised Commonsense QA with Self-Talk, Shwartz et al EMNLP 2020
One model that solves multiple commonsense tasks

<table>
<thead>
<tr>
<th>TRANSFER</th>
<th>$\alpha$NLI</th>
<th>COSMOSQA</th>
<th>HELLA SWAG</th>
<th>PIQA</th>
<th>SOCIAL IQA</th>
<th>WINOGRANDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>multitask</td>
<td>78.4</td>
<td>81.1</td>
<td>81.3</td>
<td>80.7</td>
<td>74.8</td>
<td>72.1</td>
</tr>
<tr>
<td>fine-tune</td>
<td>79.2</td>
<td>82.6</td>
<td>83.1</td>
<td>82.2</td>
<td>75.2</td>
<td>78.2</td>
</tr>
<tr>
<td>sequential</td>
<td>79.5</td>
<td>83.2</td>
<td>83.0</td>
<td>82.2</td>
<td>75.5</td>
<td>78.7</td>
</tr>
<tr>
<td>none</td>
<td>77.8</td>
<td>81.9</td>
<td>82.8</td>
<td>80.2</td>
<td>73.8</td>
<td>77.0</td>
</tr>
</tbody>
</table>
High level overview of neural LMs

Neural helps symbolic

Contextual, plug-n-play, hard to interpret

Neural methods can help with:
Knowledge acquisition
KG completion
KG correction
Fuse use KG

Future research directions:
- multitask learning with multiple KGs
- output needs to be faithful
- making model output coherent

Summary

Part-II

Symbolic helps neural

Various ways to inject CSK

CSK can help with:
Robustness
Explainability
Limited training data

Future research directions:
- topic specific paragraph KGs
- interactive learning with CSK
- multitask learning unified models

Neural helps symbolic

Neural helps symbolic

symbolic representation

Symbolic helps neural

Neural helps symbolic

Various ways to inject CSK