Carnegie Mellon University
Language Technologies Institute

An image speaks a thousand words, but can everyone listen?

On translating images for cultural relevance

Simran Khanuja
(advised by Graham Neubig, in collaboration w/ Google Research)

**disclaimer: some people may find certain content to be offensive*

Structure of the talk

The need for inclusive multi-X systems

~4-5 mins

An application: *Transcreation*

Translate images across cultures

Localize content for ads, lit/av, education, healthcare

~40 mins

Open Questions

Evaluation, Modeling, Data

~5-7 mins

The need: A case for multilingual, multimodal, multicultural systems

The need for inclusive multi-X systems

An application: *Transcreation*
Translate images across cultures
Localize content for ads, lit/av, education, healthcare

Open Questions
Evaluation, Modeling, Data

Technology and the world

The world

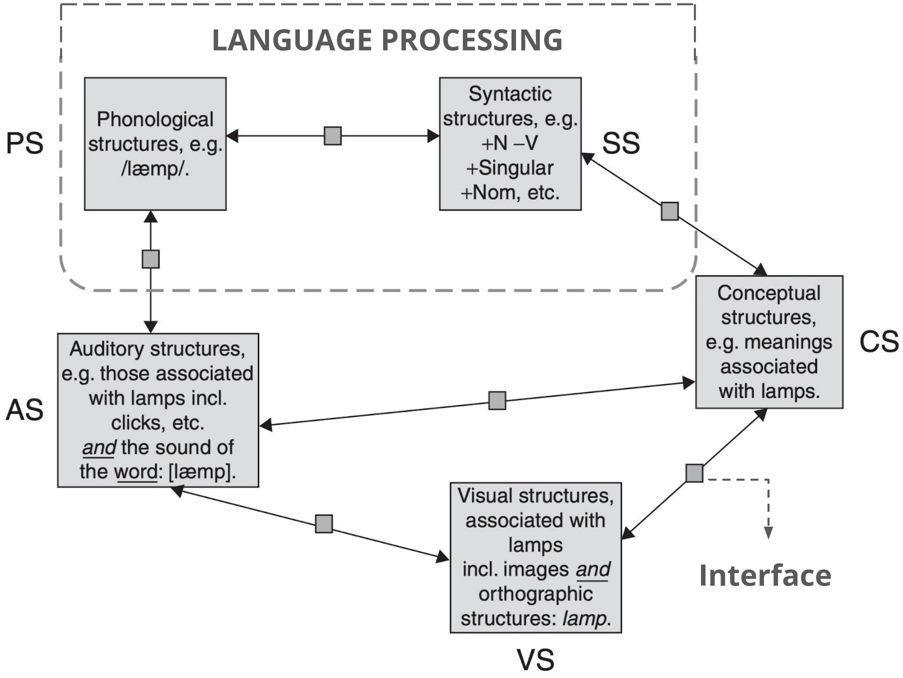
Image generated using DALL-E 3

Technology

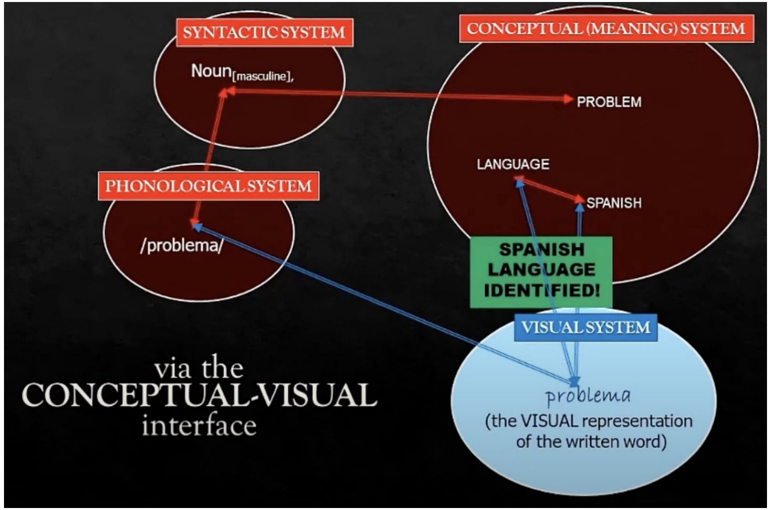
Multimodal	Multi-X	Multilingual
Flamingo	GPT4-V	NLLB
BLIP	Bard	XLM-R
ALIGN	mSLAM	GPT
CLIP	Meta MMS	mT5
LLaVa	CCLM	TuLR
LXMERT	mBLIP	mBERT
IDEFICS	MuRAL	XGLM

A neuroscience perspective:

Multimodal interfaces to a [language/mode]-neutral concept store



Example representation of "lamp" using MCF

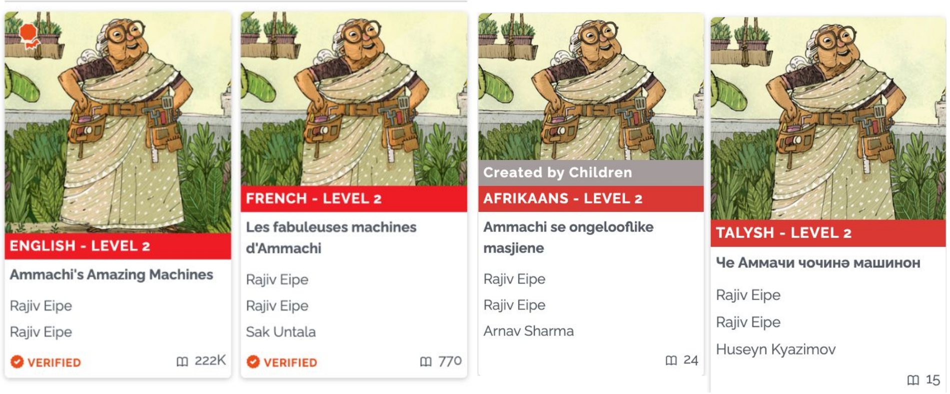


Explaining how "problema (ES)" links to the meaning of the concept "problem (EN)"

Source: The Multilingual Mind, Michael Sharwood Smith, John Truscott

***A real-world need*:**
Translating stories to different languages (and cultures?)

- Storyweaver is an organization that makes storybooks for children.
- They have stories in over 300 languages **[text]**.
- Illustrators upload independent drawings with captions **[vision]**.
- They also have read alongs with each story **[speech]**.
- They want to translate stories across borders to different languages



1. Do children refer to their grandmother as “ammachi” in all of these languages? **[text]**

2. Would a child in France, South Africa or Iran relate to this picture as that of their grandmother standing in their backyard? **[vision]**

3. What about languages that are only spoken? How do we capture regional accents,intonations **[speech]**

4. While the concept of grandmother is almost universal, what about entities like “coconut barfi” which the rest of the story is about?

The universality and non-universality of concepts

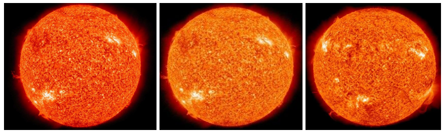
CONCEPTS

UNIVERSAL — NON- UNIVERSAL

Universal

Some concepts *look* the same across languages and cultures.

sun, gravity, black & white, fear of snakes



sun — सूरज — 太陽

Universal *(yet distinct)*

Some may universally exist but yet *look* different.

wedding, music, sports, art, religion, festivals

wedding — शादी — düğün

Non-universal

Some may not exist cross-culturally at all ...

Pilota — Shamisen — Clavie

An application: On transcreating images

The need for inclusive multi-X systems

An application: *Transcreation*

Translate images across cultures

Localize content for ads, lit/av, education, healthcare

Open Questions

Evaluation, Modeling, Data

A brief history on the definitions of **Translation**

300 AD

384 AD Jerome	*We shall try... to make not word-for-word but sense-for-sense translations* [Letter to Pammachius (384 AD): Epistulae 57.5]
1377 Ibn Khaldun	*Translation cannot occur unless the translator possesses... knowledge of the customs, manners, and mental attitudes of both nations* [The Muqaddimah (1377): Chapter 5]
1694 John Dryden	*It is a mistake to translate too literally; To imitate is one thing, and to translate another. Imitation takes the spirit of the original, but changes the dress; the translator tries to give the sense, even in different words* [Preface to Examen Poeticum (1694)]
1959 Roman Jakobson	*Interlingual equivalence.. in other words, finding the nearest natural equivalent to the semantic and syntactic unit of the source language* [On Linguistic Aspects of Translation (1959)]
1964 Eugen Nida	*Dynamic equivalence... seeks to achieve the same level of effect between receptor and text as was achieved between original author and his first audience* [Principles of Correspondence in Translating (1964)]

2000s

Machine Translation:
Tremendous progress on BLEU, and yet we make these errors

MT from 1950s to 2024

<u>Rule-based systems</u>
Large dictionaries, grammar and syntax rules

<u>Statistical systems</u>
Automatic word-alignment from large scale corpora

<u>Neural systems</u>
Deep learning, seq2seq models, transformers

<u>Large language models</u>
Large scale models trained on a plethora of data

Mistranslations today, some amusing and others expensive

<u>HSBC's "*Assume Nothing*" tagline</u>

- Mistakenly translated as "*do nothing*" in different markets.
- Bank spent $10M for replacement

<u>Pepsi</u>

- "*Come Alive With the Pepsi Generation*" arrived in China as "*Pepsi brings your relatives back from the dead*."

What is transcreation?

Defining the term

- **Trans**lation + **creation** of new content
- Why?
  - Adaptation of a message to suit the culture of the target audience
  - Preserve the intent, style, and tone of the original message
  - Evoke the same emotions

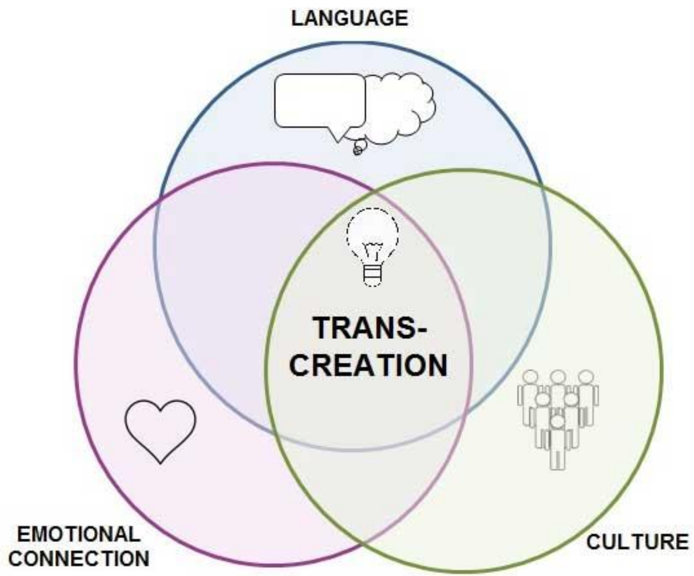


Image Source: https://digital.gov/2016/04/08/transcreation-why-do-we-need-it/

What all domains is transcreation prevalent today?

Healthcare

Design interventions that resonate with the community experiencing health disparities

Advertisements

Think Global, Act Local

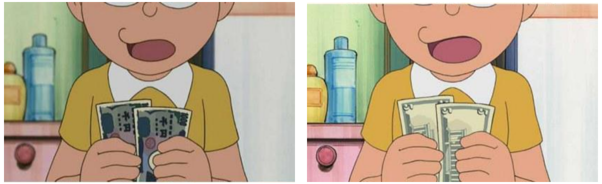
Global brands usually need to localize ads

Education

Teaching counting

(*left*: US; *right*: India)

Literature/Audiovisual translation



Doraemon: change Yen to USD

Storyweaver

Peter Parker → Pavitr Prabhakar
Mary Jane → Meera Jain
Aunt May → Auntie Maya
Harry Osborne → Hari Oberoi

Spider-man India

Our Goal

To assess the capabilities of state-of-the-art generative AI technology to aid the process of translating visual content across cultures

in the words of a friend ...

If the same perfect storm of artistic coincidences had happened in a different culture, in a different time -- what would it have looked like?

[Pipeline 1]: InstructPix2Pix

Image editing using natural language instructions

Generate text edits

Input Caption: *"photograph of a girl riding a horse"*
Instruction: *"have her ride a dragon"*

→ GPT-3 (finetuned) →

Edited Caption: *"photograph of a girl riding a dragon"*

Generate paired images

Input Caption: *"photograph of a girl riding a horse"*
Edited Caption: *"photograph of a girl riding a dragon*

→ Stable Diffusion + Prompt2Prompt →

Generate training examples

"have her ride a dragon"

"Color the cars pink"

"Make it lit by fireworks"

"convert to brick"

...

[Pipeline 1]: InstructPix2Pix

Image editing using natural language instructions

Advantages

Why InstructPix2Pix over other image-editing models?

1. *Abstract NL instructions →*
   prompt-to-prompt for comparison
2. *No extra input →*
   like captions, segmentation masks
3. *Very fast →*
   Performs edit in forward pass without need for inversion
4. *Widely used →*
   Max. downloads on HF

Results

Instruction
Make this image culturally relevant to Japan

Visualization Link (Japan)

Does not retain semantic coherence →
inserts objects out of context, based on colors/shapes

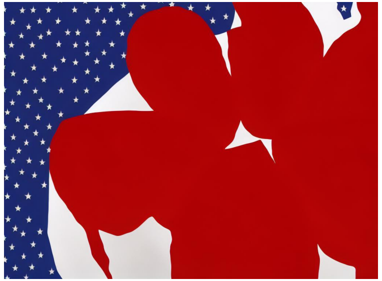
Exhibits strong color bias →
like red/black for Japan, brown/black for Nigeria

Changes people in deterministic ways →
Open: Is this a good or a bad thing? Where do we draw the line b/w relatability v/s offensiveness?

Lacks understanding of cultural entities→
edits entities specific to a culture, potential to seriously harm sentiments

[Pipeline 1]: InstructPix2Pix

Quiz!



[Pipeline 2] Caption → Edit for cultural relevance → Image Edit

BLIP → GPT3.5 → PlugnPlay

Methodology

Step 1: Caption the image using BLIP

a field of cotton plants

Step 2: Edit the caption for cultural relevance using GPT-3.5

Prompt

```
Edit the input text, such that it
is culturally relevant to Japan.
Keep the output text of a similar
length as the input text. If it is
already culturally relevant to
Japan, no need to make any edits.
The output text must be in English
only.
Input: a field of cotton plants
Output:
```

Output

```
a rice paddy field
```

Step 3: Edit the original image using o/p from Step-2

a rice paddy field

[Pipeline 2] Caption → Edit for cultural relevance → Image Edit

BLIP → GPT3.5 → PlugnPlay

Error Types (target: India)

Issues with captioning

a man in white sari standing in a field

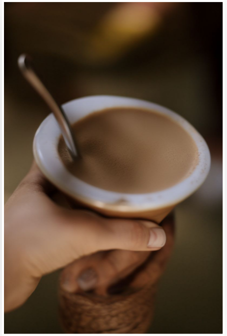
Issues with LLM editing

A bowl of ramen with meat and vegetables

A bowl of ramen with chicken and vegetables

Issues with image editing due to preservation of spatial layout

a person holding a cup of green tea



a person holding a cup of chai

[Pipeline 3] Caption → Edit for cultural relevance → Retrieval

BLIP → GPT3.5 → LAION (Country-specific)

Methodology

Step 1: Caption the image using BLIP

a field of cotton plants

Step 2: Edit the caption for cultural relevance using GPT-3.5

Prompt

```
Edit the input text, such that it
is culturally relevant to Japan.
Keep the output text of a similar
length as the input text. If it is
already culturally relevant to
Japan, no need to make any edits.
The output text must be in English
only.
Input: a field of cotton plants
Output:
```

Output

```
a rice paddy field
```

Step 3: Retrieve most similar image to text o/p in Step-2 from LAION-JP

(filter URLs containing ".jp" in the domain)

a rice paddy field

[Pipeline 3] Caption → Edit for cultural relevance → Retrieval

BLIP → GPT3.5 → LAION (Country-specific)

Error Types (target: India)

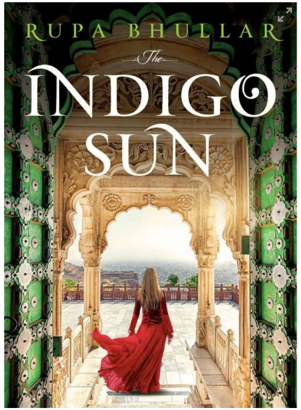
Does not preserve spatial layout

a person holding a cup of green tea

a person holding a cup of chai

Collision, retrieves irrelevant o/ps

a sunflower is standing in front of a blue sky



a sunflower is standing in front of a blue sky

Offensive outputs

This pipeline is as good as the database of images it can retrieve from

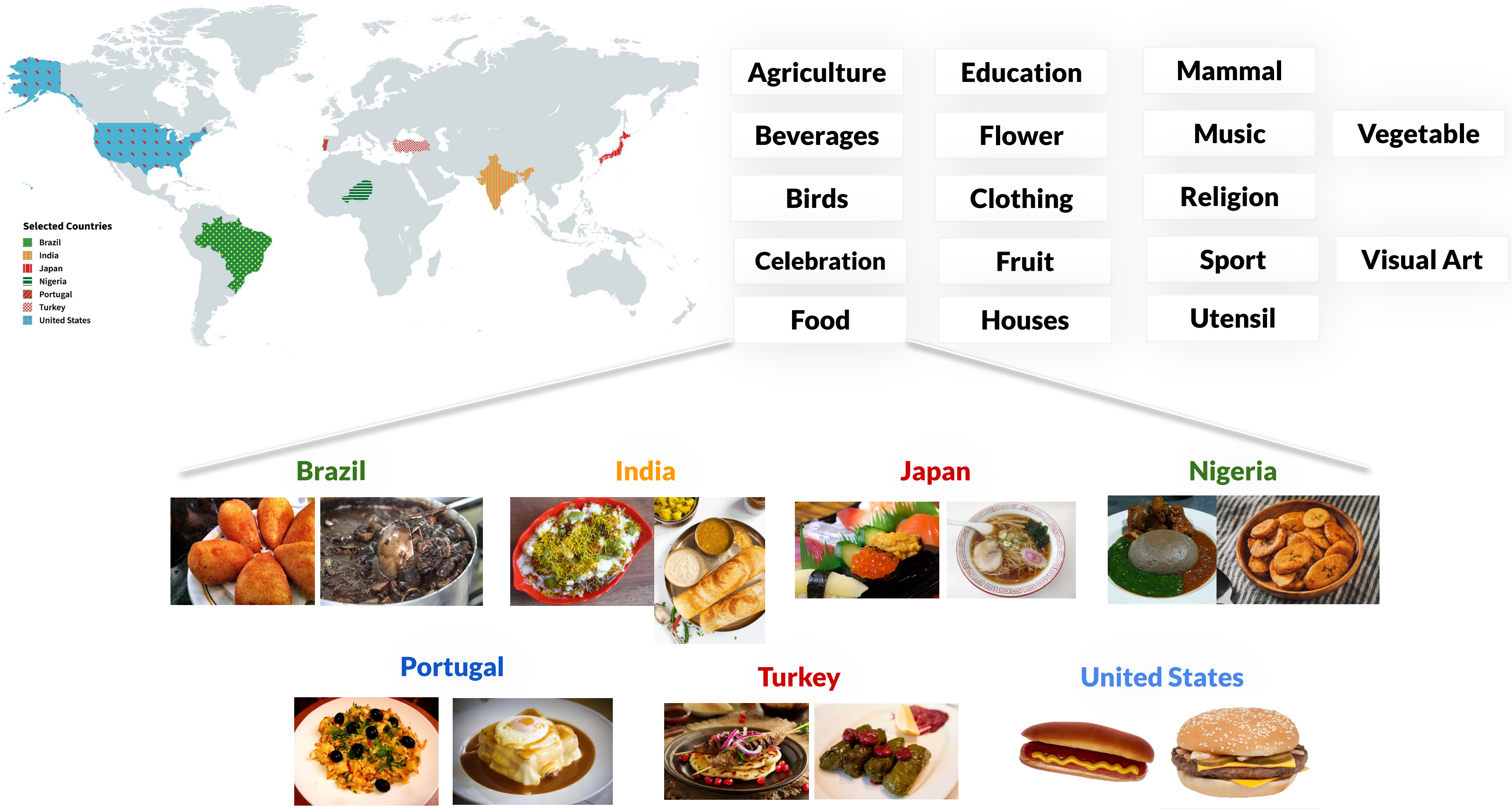
It can sometimes retrieve very offensive images due to collision issues as highlighted

[Part-1] Evaluation : Concept / Object Level

Cultural concepts selected from 7 countries across 17 categories

Data Collection Methodology

1. Selected 7 geographically diverse countries
   a. Brazil, Japan, India, Nigeria, Portugal, Turkey, United States
2. Listed 17 semantic categories from the Inter-continental Dictionary Series
   a. Agriculture, birds, beverages, mammals, food, education, religion, music, visual arts ...
3. Hired annotators to list 5 concepts in each category such that they are:
   a. commonly seen or representative of the speaking population of your country
   b. ideally, to be physical and concrete
4. Collected ~600 images for each country
   a. Brazil, Japan, India, Nigeria, Portugal, Turkey, United States
5. Results from all pipelines (randomized)
   a. Brazil, Japan, India, Nigeria, Portugal, Turkey, United States


Selected Countries
Brazil
India
Japan
Nigeria
Portugal
Turkey
United States
Agriculture
Education
Mammal
Beverages
Flower
Music
Vegetable
Birds
Clothing
Religion
Celebration
Fruit
Sport
Visual Art
Food
Houses
Utensil
Brazil
India
Japan
Nigeria
Portugal
Turkey
United States

[Part-2] Evaluation : Application-oriented

Task-oriented images for education and literature

Data Collection / Pipelines

1. Curated ~70 images for education and ~40 for literature
   a. Education: K5 Learning (US) & NCERT (India)
   b. Literature: Storyweaver (India)

2. Used worksheet task / story title to generate appropriate captions / LLM edits (Japan pipeline-2)

Ammachi's amazing machines

InstructBLIP

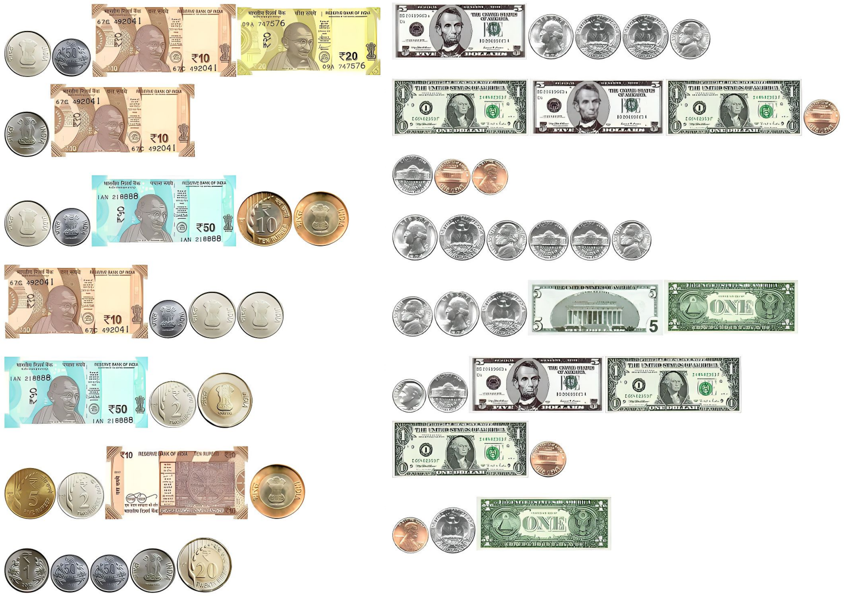
an old woman in a sari is holding a tool belt

GPT 3.5

An elderly woman in a kimono is carrying a tool box

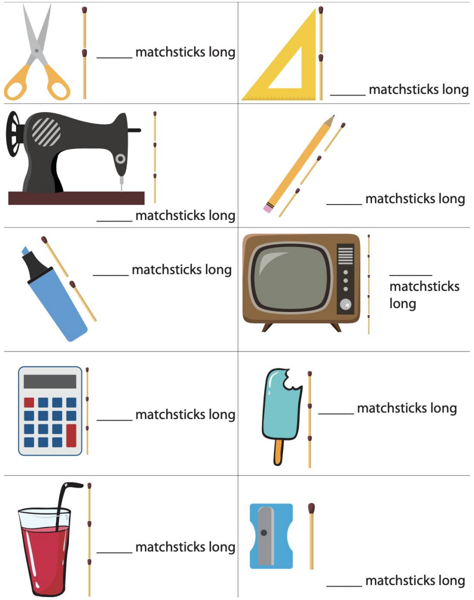
PlugnPlay

Evaluation [Part-2]: Application-oriented
Education Examples



Teaching addition with a currency worksheet *(left: India; right: US)*

Halloween-themed w/sheet teaching counting



Teaching how to measure with matchsticks

Evaluation: Why the two-part evaluation?

Discussion

1. Eventual goal is to apply it to part-2
2. Real world images are complex scenes comprised of multiple objects
3. Part-1 goals are to:
    a. Provide a simpler dataset with one image per concept/object
    b. Diversity helps discern performance across varied categories
    c. Hope is for models to make progress towards part-2 using part-1 (compositionality)

Human Evaluation: Questions asked

ID	Question	Property	Applications	Performance
Concept Dataset				
C0	Is there any visual change in the generated image compared to the original image?	`visual-change`	None (*helps filter non-edits*)	`e2e-instruct` `cap-edit` `cap-retrieve`
C1	Is the generated image from the same semantic category as the original image?	`semantic-equivalence`	AV (Zootopia); Education	`e2e-instruct` `cap-edit` `cap-retrieve`
C2	Does the generated image maintain spatial layout of the original image?	`spatial-layout`	AV (Doraemon, Inside Out)	`e2e-instruct` `cap-edit` `cap-retrieve`
C3	Does the image seem like it came from your country/ is representative of your culture?	`culture-concept`	AV, Education, Ads	`e2e-instruct` `cap-edit` `cap-retrieve`
C4	Does the generated image reflect naturally occurring scenes/objects?	`naturalness`	Ads (Ferrero Rocher)	`e2e-instruct` `cap-edit` `cap-retrieve`
C5	Is this image offensive to you, or is likely offensive to someone from your culture?	`offensiveness`	All	`e2e-instruct` `cap-edit` `cap-retrieve`
-	For edited images, is the change meaningful (C1) and culturally relevant (C3)?	`meaningful-edit`	All	`e2e-instruct` `cap-edit` `cap-retrieve`
Application Dataset				
E1	Can the generated image be used to teach the concept of the worksheet?	`education-task`	Education	`e2e-instruct` `cap-edit` `cap-retrieve`
S1	Would the generated image match the title of the story in a children's storybook?	`story-title`	AV, Literature	`e2e-instruct` `cap-edit` `cap-retrieve`
E/S2	Does the image seem like it came from your country/is representative of your culture?	`culture-application`	All	`e2e-instruct` `cap-edit` `cap-retrieve`
-	For edited images, is the change meaningful (E/S1) and culturally relevant (E/S2)?	`meaningful-edit`	All	`e2e-instruct` `cap-edit` `cap-retrieve`

[Part-1] Human Evaluation: only 6% translations successful for some

Figure 6: *Human ratings for the concept dataset*: Our primary goal is to test whether the edited image belongs to the same universal category as the original image (**C1**) and whether it increases cultural relevance (**C3**). We plot the count of images that can do both above (**C1+C3**), and observe that the best pipeline's performance ranges between 6% (Nigeria) to 30% (India).

[Part-2] Human Evaluation: no translations successful for some

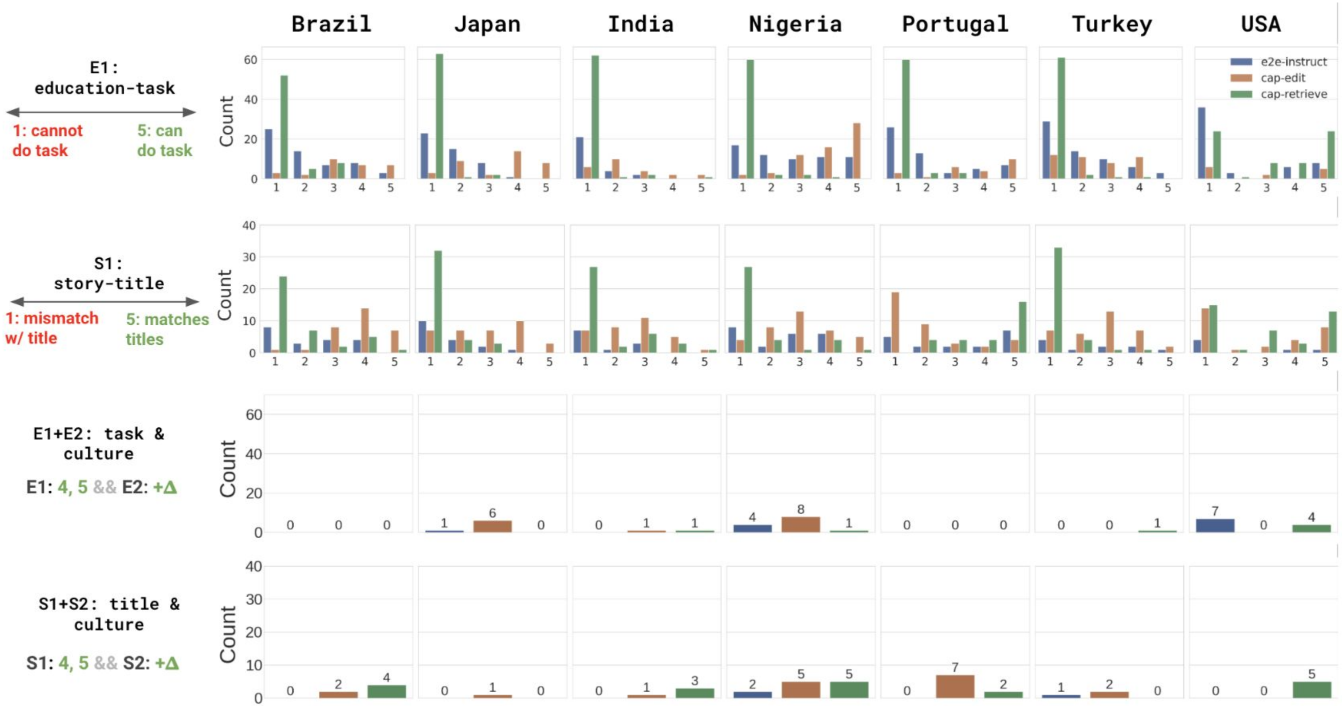


Figure 7: *Human ratings for the application dataset*: Our goal is to test whether the edited image can be used for the application as before (**E/S1**), and whether it increases cultural relevance (**E/S2**). We plot the count of images that can do both above (**E/S1+E/S2**), and observe that even the best pipeline cannot translate any image successfully in some cases, like for Brazil and Portugal in education.

Open Questions in evaluation, modeling and data

The need for inclusive multi-X systems

An application: *Transcreation*
Translate images across cultures
Localize content for ads,
lit/av, education,
healthcare

Open Questions
Evaluation, Modeling, Data

Food for Thought

Evaluation

1. Do models incorporate diversity in representation?
    a. Initial explorations suggest otherwise
    b. [Open] How do you evaluate diversity?
        i. [Open] Can you account for individual preferences?
2. What is the tradeoff between diversity v/s stereotyping/bias?
    a. [Open] Can models produce diverse outputs with diff. initializations/conditioning?
3. How does one decide what is most culturally appropriate to a user?
    a. [Open] Is it right to discern culture based on language input?
        i. English is ubiquitous
        BUT, also
        ii. Language has evolved within a culture and holds key information about it
    b. How do you account for individual experiences, example, the children of immigrants?

Food for Thought

Data and Modeling

1. Do models have a world view of concepts specific to every culture?
    a. Probably not and may never will
        i. Not everything is present digitally
        ii. Cultures and concepts are constantly changing
2. How can we make models adept at keeping up with evolving concepts and cultures?
3. How can we incorporate cultures of communities that are not present digitally, into our models?
4. Learning from multilingual, multimodal data is very hard
    a. What kind of an architecture should such a system have?
        i. Maybe the MCF framework can help?
    b. How do we design optimal learning objectives?
5. How do we obtain data annotations at a cultural level? How do we make a distinction between semantic drifts for the same concepts across multiple cultures?

Thanks! Questions?

skhanuja@andrew.cmu.edu