Vision + Language

Guest Lecture by Andrea Burns Spring 2021

Content adapted from Bryan A. Plummer

Lecture Outline

- What is “vision and language”? + example tasks
- Visual representations (review)
- Language representations
- Example task architectures
- Recent work

What are vision-language problems?

Input → output might be...

A. Image → text
B. Text → image
C. Image + text → text

Examples...

A. Image Captioning
B. Sentence to Image Retrieval
C. Visual Question Answering

Image Captioning

A woman in a white shirt and denim overalls is walking six dogs in the park.

Generating natural language sentences given the input image

Sentence to Image Retrieval

A woman in a white shirt and denim overalls is walking six dogs in the park.

Retrieving an image given the input natural language sentence

Sentence to Image Retrieval

A woman in a white shirt and denim overalls is walking six dogs in the park.

Retrieving an image given the input natural language sentence

Visual Question Answering

\+ How many dogs are being walked? → Six

Generating an answer to a visual question, often framed as a classification problem

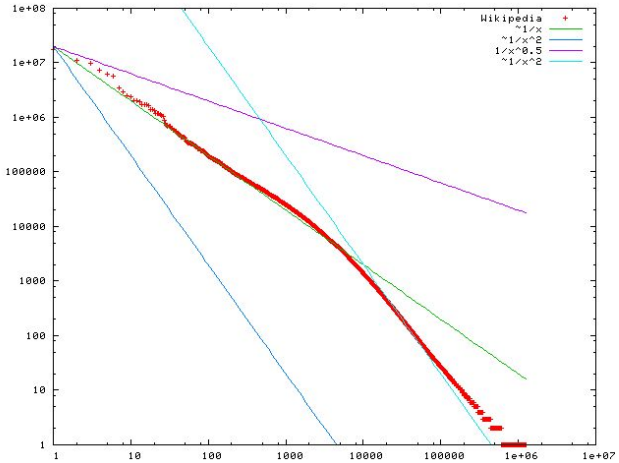
How are vision-language tasks different?

Zipf's Law

The frequency of a word is inversely proportional to its rank in a frequency table

Few words occur very often, and they aren't the most semantically rich

Word Frequency on Wikipedia



Word Frequency

Word Rank

How would you describe this image?

- An orchestra performing in a concert hall.
- The Dublin Symphony Orchestra playing in an ornate theater.
- An auditorium with many musicians playing and an audience around them.

How would you describe this image region?

- A man
- A bearded man
- A person with sunglasses and a hat on
- A man with beard, tinted shades, a hat and a leather jacket with a red t-shirt underneath

Need for Abstract or Commonsense Reasoning

Two siblings are walking on rocks across a river

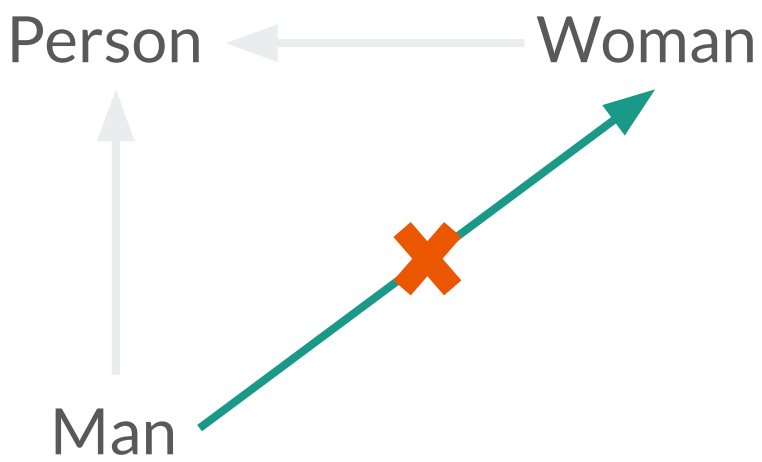
One pair of siblings

They must be moving

People can't walk on water

Rocks must be fixed in river

Non-Transitive Semantic Relationships



Evaluation Metrics

- Difficult to capture how “good” generated language is
- BLEU (Bilingual Evaluation Understudy)
  - Take unigram, bigram, or trigram etc. of generated sentence and check if it occurs in ground truth
  - Divide by # of words in generated sentence & cap frequency of n-gram
    - “I ate three hazelnuts” vs. “three three three three”
  - Sum over score from each n-gram

Ground Truth

I ate three hazelnuts.

Generated

I had three nuts.

I, had, three, nuts	unigrams
I had, had three, three nuts	bigrams
I had three, had three nuts	trigrams

¼ + 0 + ¼ + 0 = 0.5

Evaluation Metrics

- BLEU - what are the problems with this?

Ground Truth

I ate three hazelnuts.

Generated

I had three nuts.

I	had	three	nuts	unigrams
I had	had three	three nuts		bigrams
I had three	had three nuts			trigrams

- Doesn't consider word meaning
- Doesn't consider relative word importance
- Unigrams do not consider word order

Evaluation Metrics

- Difficult to capture how “good” generated language is
- CIDEr (Consensus based Image Description Evaluation)
  - **CIDEr(a, b) = *cos sim*(g^n(a), g^n(b))**
  - Where g^n(x) = vector formed from TF-IDF score of each n-gram

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Evaluation Metrics

- Difficult to capture how “good” generated language is
- CIDEr (Consensus based Image Description Evaluation)
  - **CIDEr(a, b) = *cos sim*(g^n(a), g^n(b))**
  - Where g^n(x) = vector formed from TF-IDF score of each n-gram

Pros? Cons?

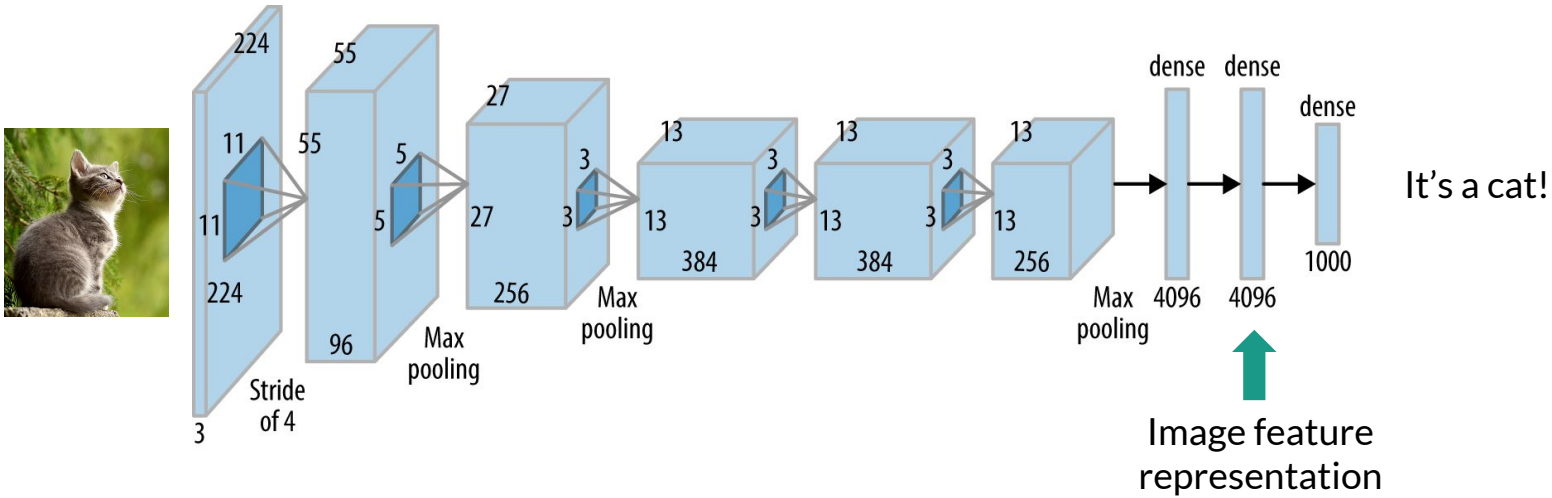
- Has relative importance of words now
- Still doesn’t really take meaning into account

Challenges

- Words can relate to objects, their attributes, relationships between entities, or require commonsense reasoning
- Non-transitive semantic relationships
- Very sparsely labeled + data difficult to collect
- Difficult to evaluate quality of generated text
- Many words seen at test time may not be present in the training data
- Word ordering may change what kinds of images you would expect to see

Representation Learning

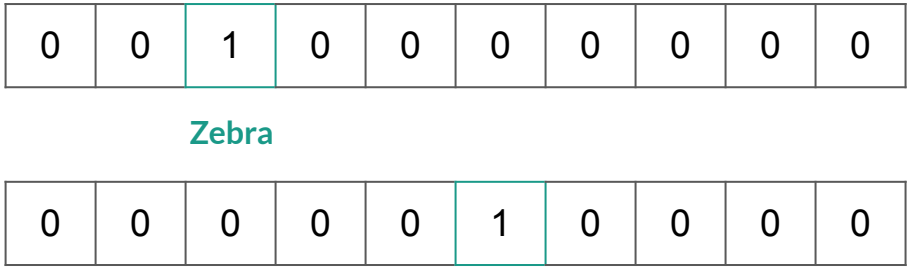
Convolutional Neural Networks (CNN)



Krizhevsky et al. "ImageNet classification with deep convolutional neural networks." NeurIPS, 2012.

Language Features

- What is a simple way to represent words?
  - Treat each word as a class (one-hot vector representation)



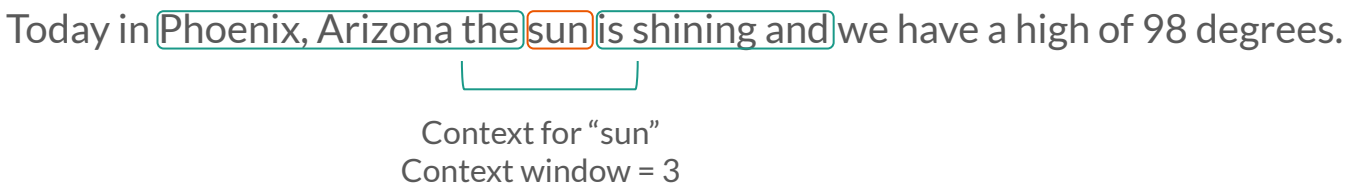
How many words can we represent with this?

What are downsides to one-hot vectors?

- Size of vector increases as vocabulary size increases
- No semantics are contained in one-hot vectors
  - They’re all orthogonal!

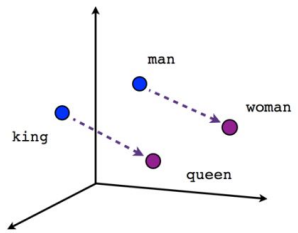
Semantically Rich, Dense Embeddings

- How can we obtain dense, fixed size language features?
  - “A man is known by the company he keeps” - Aesop
  - Concept in NLP community is “context”

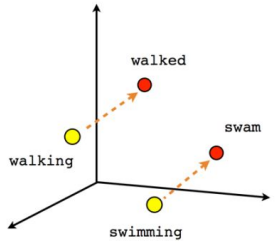


Word2Vec

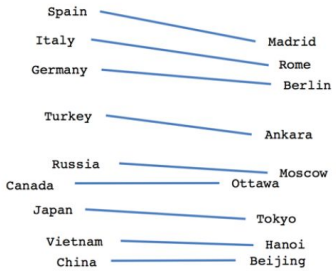
- This is how Word2Vec was trained
  - Given context words, predict current word (CBOW)
  - Given current word, predict context words (SkipGram)
  - Start with one-hot or randomly init vectors, trained neural network with this classification objective
  - Can be trained completely unsupervised!!



Male-Female



Verb tense

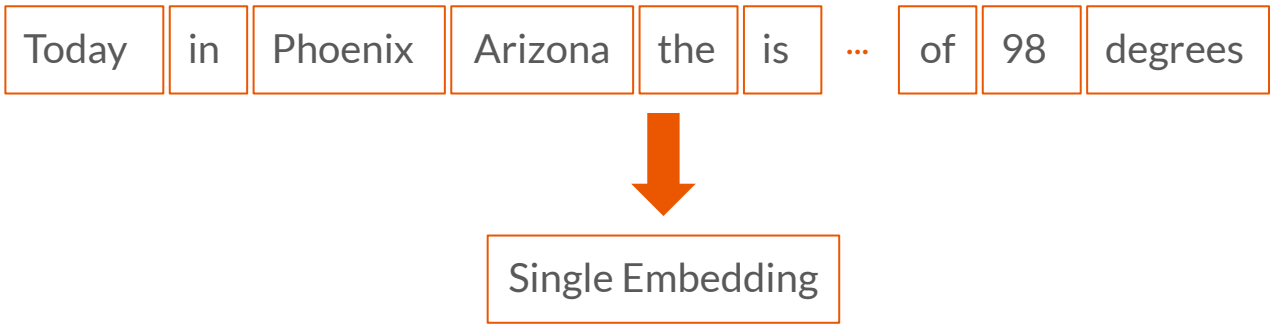


Country-Capital

So now what?

- If our language input is larger than single words (*i.e.*, sentences, paragraphs, etc.) we have to decide how to aggregate word embeddings

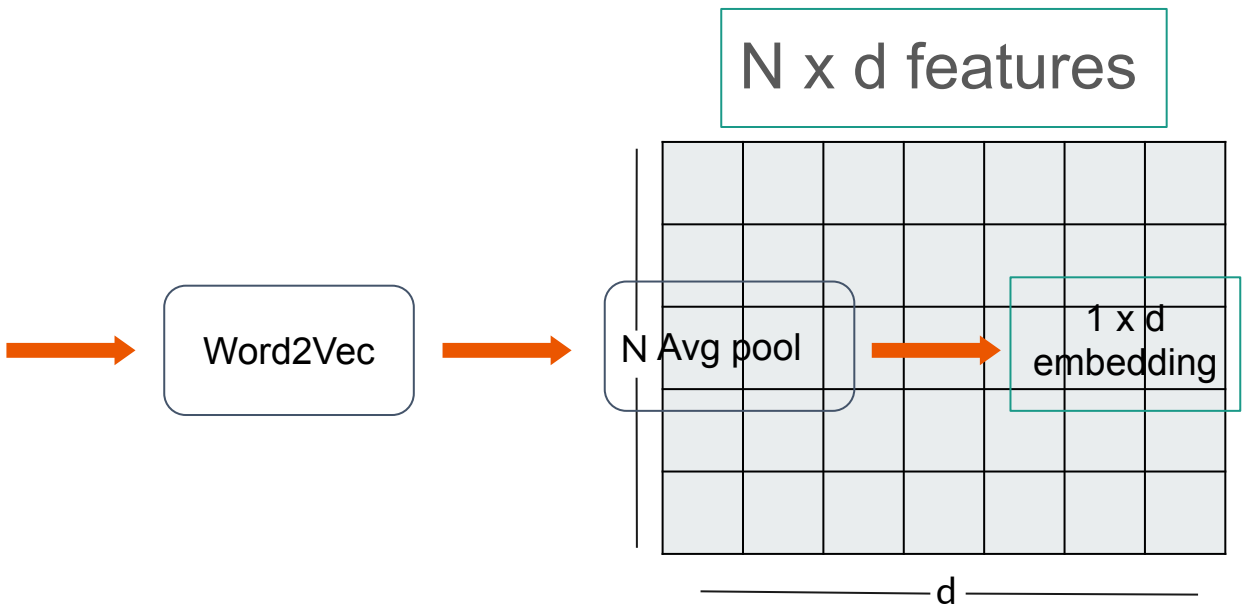
Today in Phoenix, Arizona the sun is shining and we have a high of 98 degrees.



Average Embedding

N words

Today in Phoenix, Arizona the sun is shining and we have a high of 98 degrees.

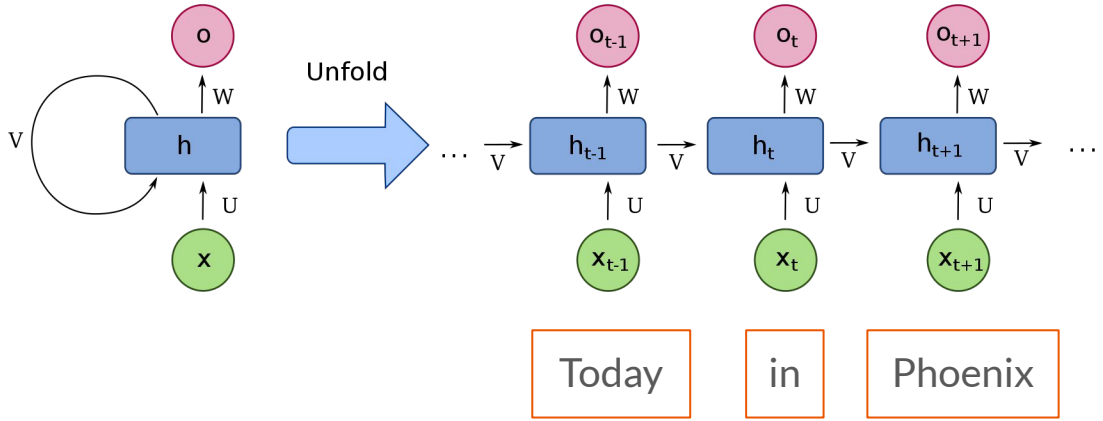


Average Embedding

- Pros?
  - Simple + efficient
    - No parameters needed!
  - Often works well and doesn't reduce performance
- Cons?
  - Loses information like word ordering
  - Can't be used in generative tasks
    - *E.g.* image captioning

LSTM Representation

- Take last hidden state as sequential representation



LSTM

- Pros?
  - Better at representing sequential data
- Cons?
  - Adds additional parameters to the network
  - Can be relatively slow for large layers

Common Task Architectures & Objectives

Image-Sentence Retrieval

Neural Net

Feature Set A

"An orchestra performing in a concert hall."

Neural Net

Feature Set B

Shared Semantic Space

Image-Sentence Retrieval

Common loss is triplet loss

$L_{triplet}(x, y^+, y^-) = max(0, m + d(x, y^+) - d(x, y^-))$

"Two women strolling on a boardwalk by the beach."

$d(x, y^-)$

$d(x, y^+)$

"An orchestra performing in a concert hall."

Want positives closer in embedding space, but allow for margin so there isn't strict constraint

Minimizing triplet loss should result in:
$d(x, y^+)\downarrow$
$d(x, y^-)\uparrow$

Image-Sentence Retrieval

Common loss is triplet loss

$$L_{triplet}(x, y^+, y^-) = max(0, m + d(x, y^+) - d(x, y^-))$$

Can use both images and sentences as anchors:

$$L_{mm} = L_{triplet}(I, S^+, S^-) + L_{triplet}(S, I^+, I^-)$$

"An orchestra performing in a concert hall."

"The Dublin Symphony Orchestra playing in an ornate theater."

"An auditorium with many musicians playing and an audience around them."

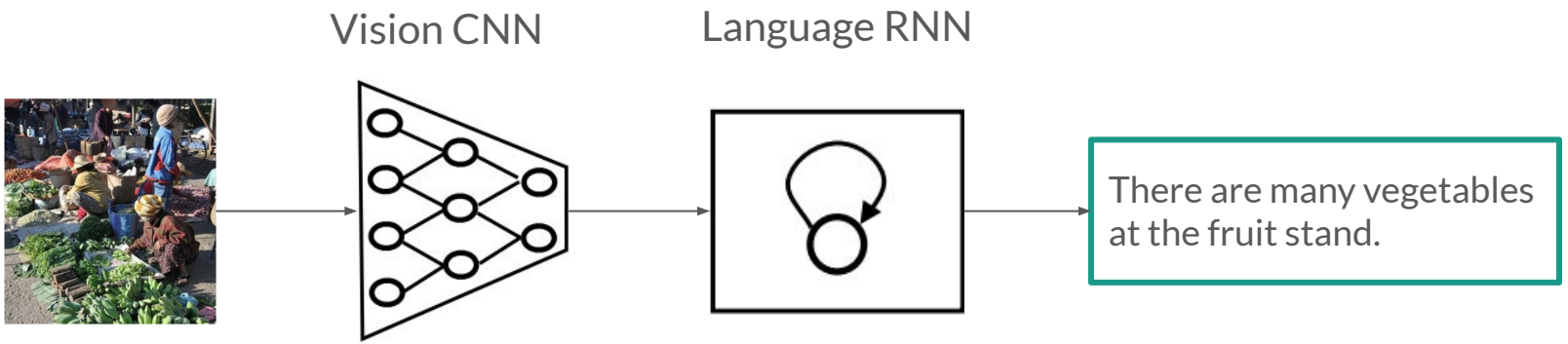
One image usually has 5 paired sentences

Image Captioning

Vinyals et al., Show and Tell: A Neural Image Caption Generator, CVPR, 2015.

What can we borrow from NLP literature?

- Inspired by Machine Translation: Encoder → Decoder framework



LSTM inputs

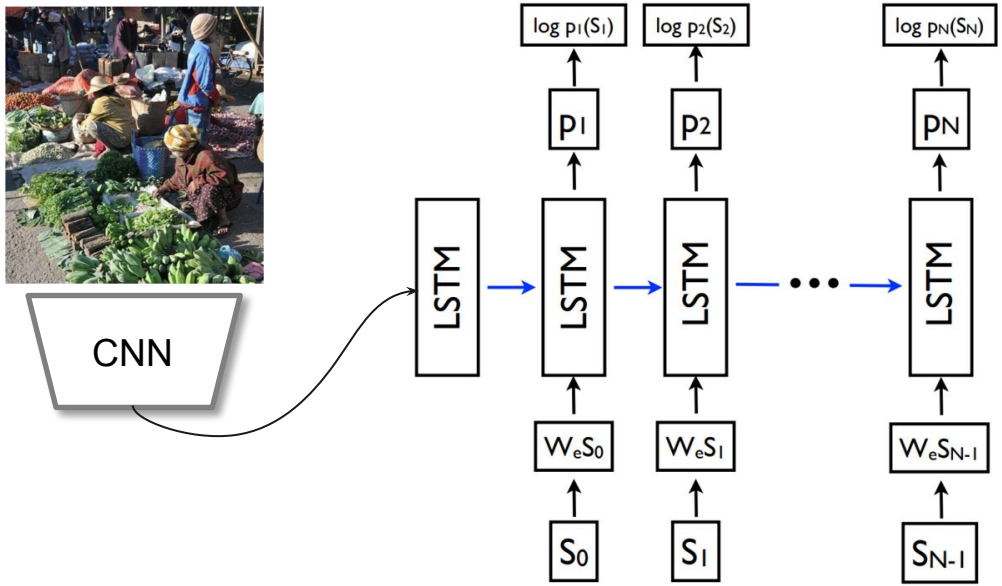


image source: Show and Tell: A Neural Image Caption Generator, CVPR, 2015

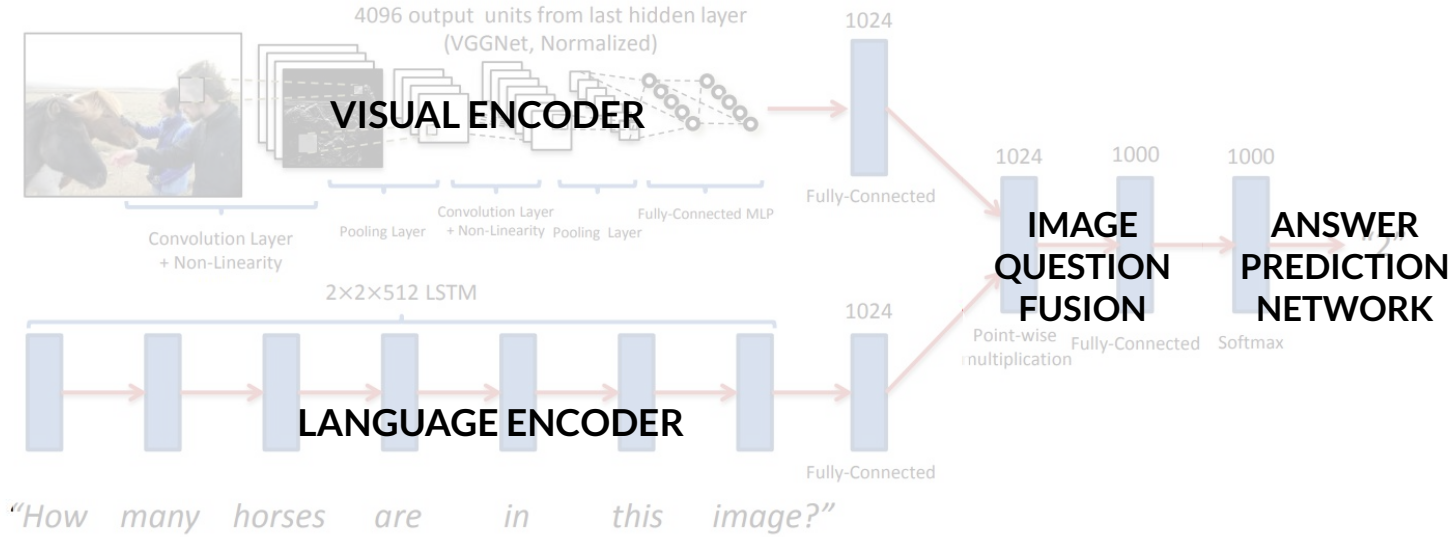
Image Captioning

What is our training objective?

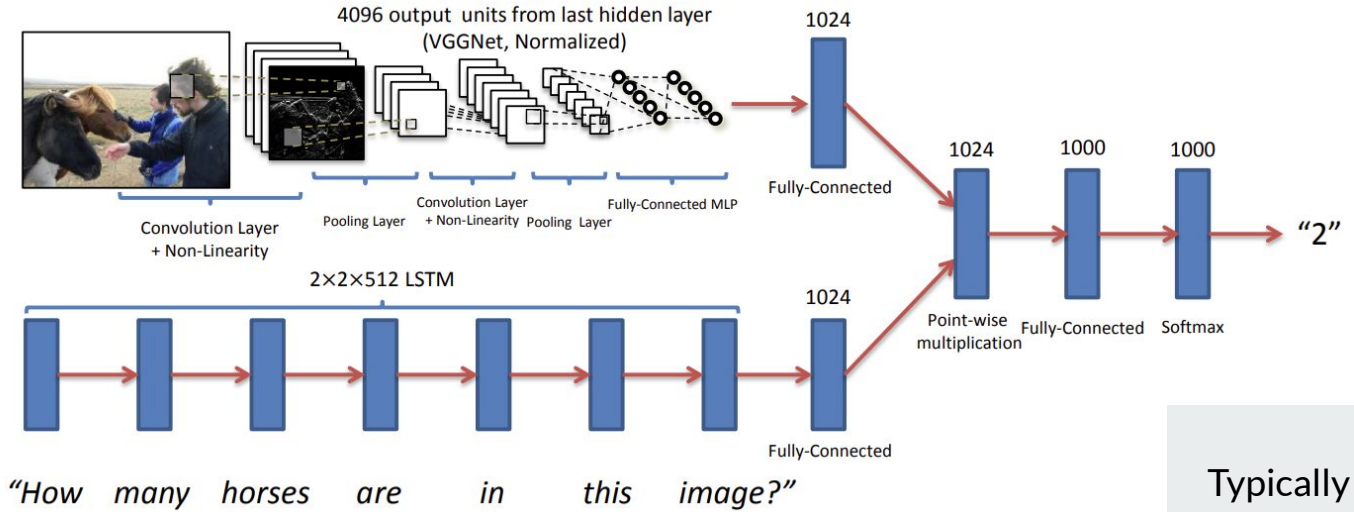
- Maximize the likelihood of a caption

$$\text{Negative Log Likelihood (NLL)} = -\sum \textit{log}\, p_t(s_t)$$

Visual Question Answering



Visual Question Answering



Typically trained with cross entropy

Recent work

Representation Learning

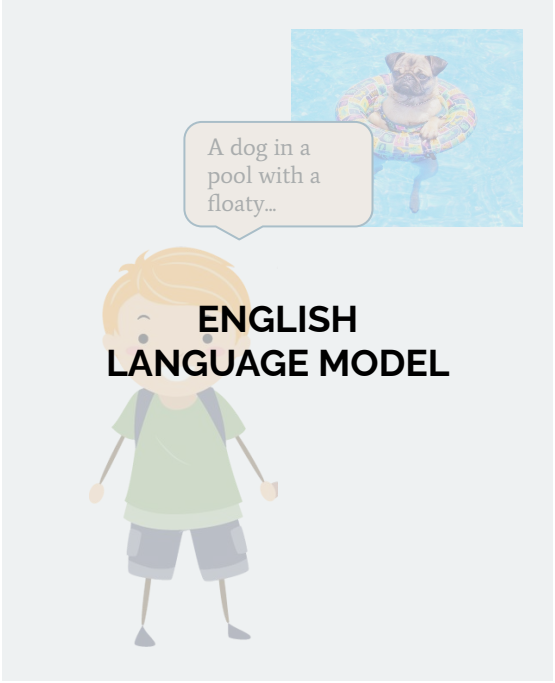
- It’s a large field with many research directions
  - Joint vision-language features
  - Hierarchical/structured representations
  - Multilingual extensions
  - Removing language biases

Learning to Scale Multilingual Representations for Vision-Language Tasks

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, Bryan A. Plummer

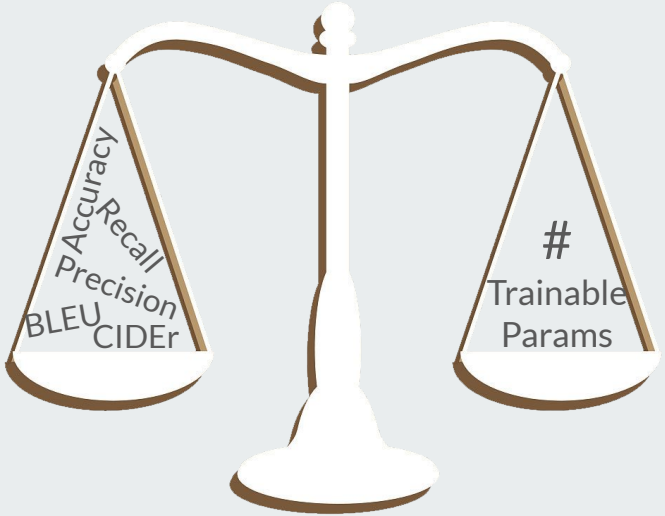
- Extending English-only vision-language tasks, what challenges arise?
  Why do we want to do this?
  - Make downstream applications accessible to all (from any part of the world, thousands of languages)
  - Representing many languages equally well
  - Size of language model as we add more languages

Existing Multilingual Vision-Language Work



What about a single language model?

Performance / Scalability



Multilingual Vision-Language Models

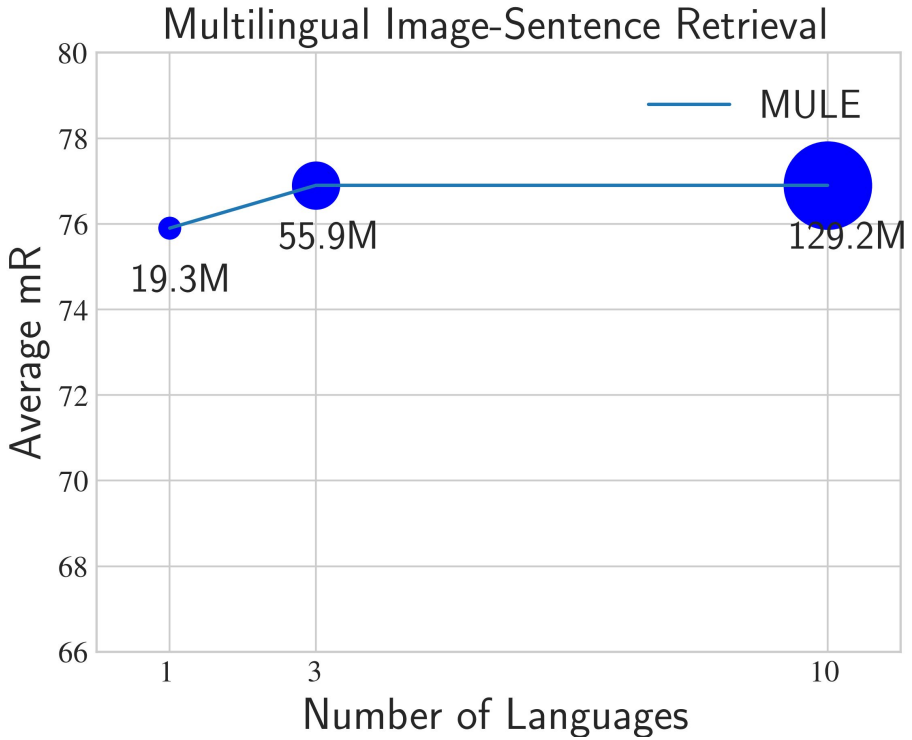
Do we have to choose?

Goal: Build a scalable model that doesn't come at the cost of performance

MULE

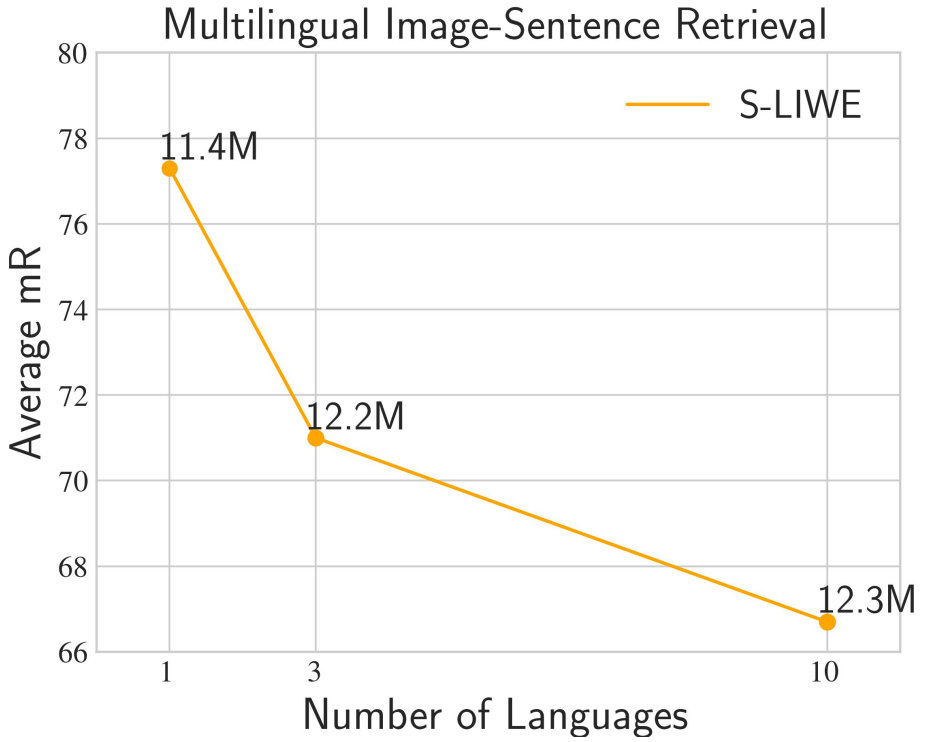
MULE: Multimodal Universal Language Embedding, AAAI 2020

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, Bryan A. Plummer



LIWE

Language-Agnostic Visual-Semantic Embeddings, ICCV 2019
Jonatas Wehrmann, Douglas M. Souza, Mauricio A. Lopes, Rodrigo C. Barros

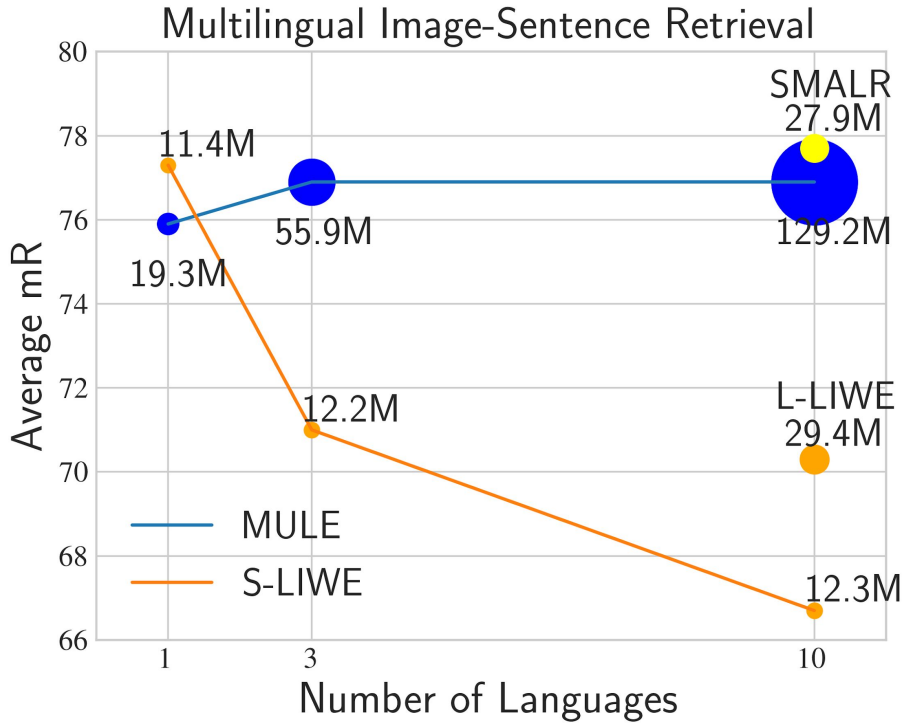


SMALR

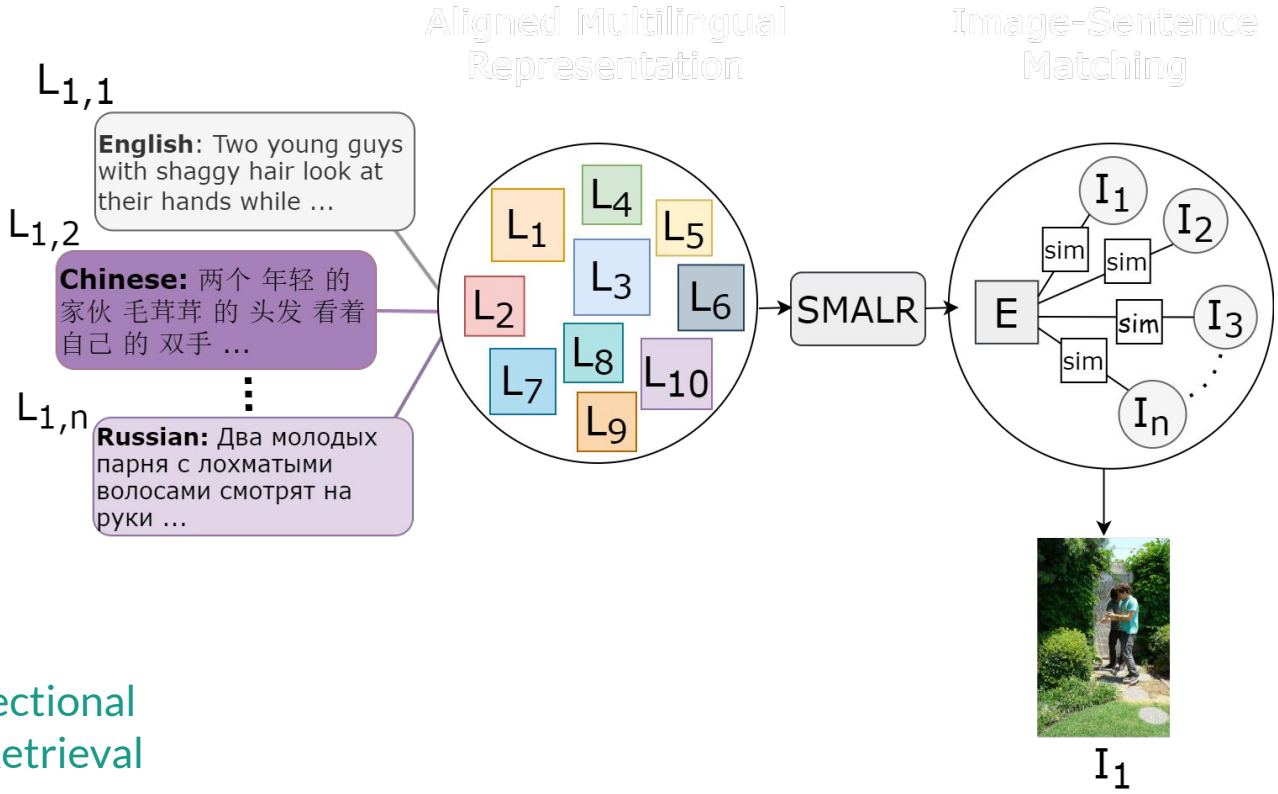
Our proposed model for multilingual representation learning

Learning to Scale Multilingual Representations for Vision-Language Tasks, ECCV 2020

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, Bryan A. Plummer

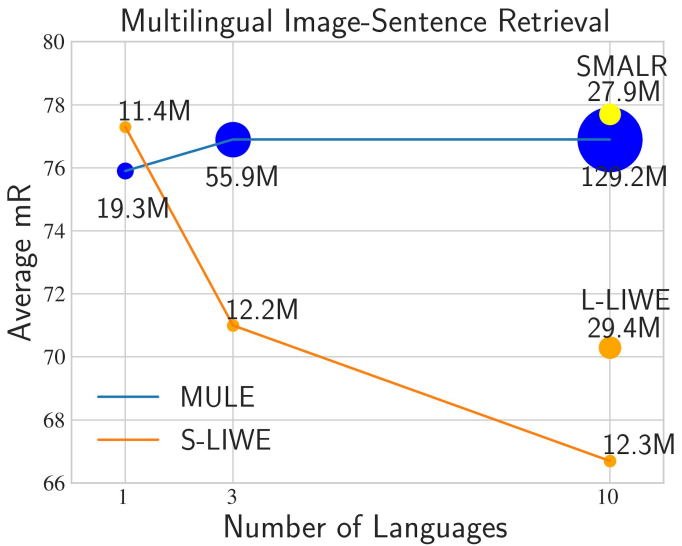


Task



Multilingual Bidirectional
Image-Sentence Retrieval

Our Model: SMALR

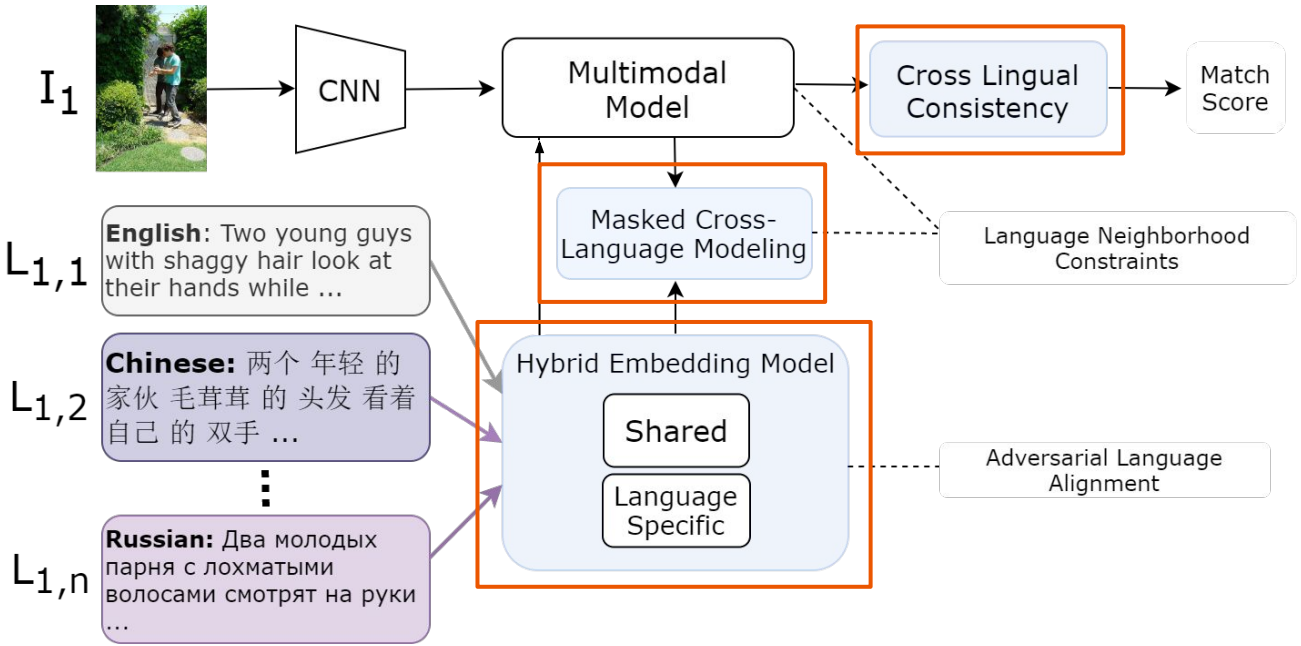


We present **SMALR**
Scalable Multilingual Aligned Language Representation

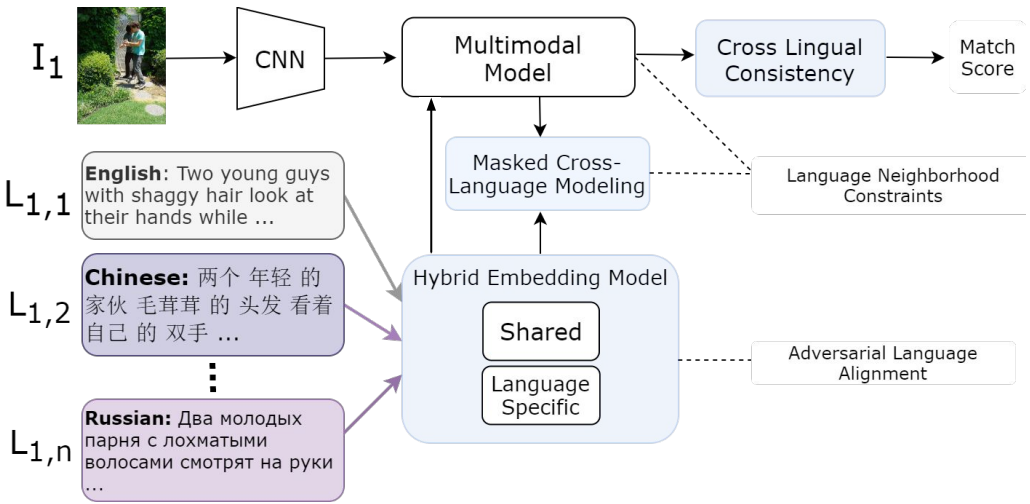
Evaluated on **10 languages**

- English
- German
- French
- Czech
- Chinese
- Japanese
- Korean
- Russian
- Afrikaans
- Arabic

Model


I_1
CNN
Multimodal Model
Cross Lingual Consistency
Match Score
Masked Cross-Language Modeling
Language Neighborhood Constraints
$L_{1,1}$
English: Two young guys with shaggy hair look at their hands while ...
$L_{1,2}$
Chinese: 两个 年轻 的 家伙 毛茸茸 的 头发 看着 自己 的 双手 ...
$L_{1,n}$
Russian: Два молодых парня с лохматыми волосами смотрят на руки ...
Hybrid Embedding Model
Shared
Language Specific
Adversarial Language Alignment

Model



SMALR is trained end-to-end with 4 loss terms

$$L_{SMALR} = L_{mm} + \lambda_2 L_{mask} + \lambda_3 L_{adv} + \lambda_4 L_{nc}$$

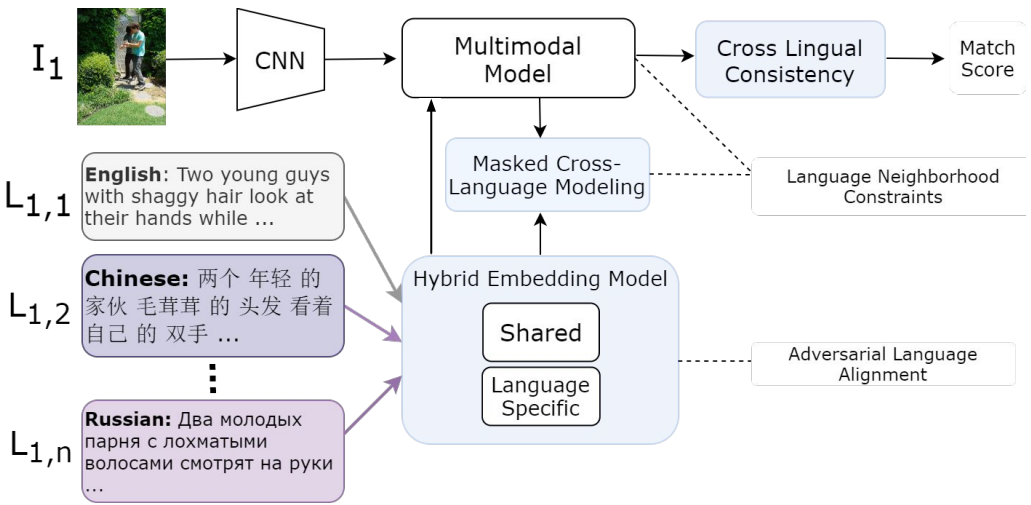
L_{mm} = Multimodal model loss

L_{mask} = Masked Cross Language Modeling Loss

L_{adv} = Adversarial Language Alignment Loss

L_{nc} = Neighborhood Constraint Loss

Model

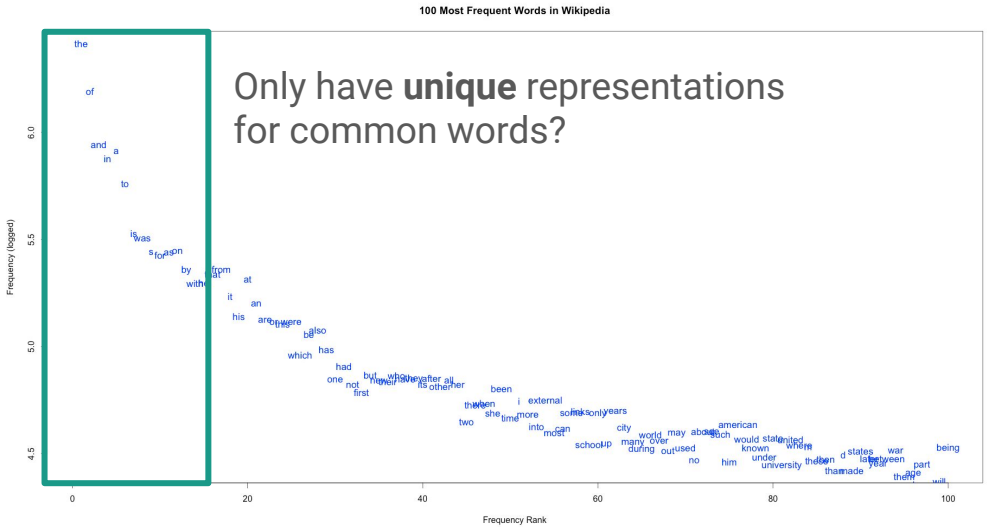


$$L_{SMALR} = L_{mm} + \lambda_2 L_{mask} + \lambda_3 L_{adv} + \lambda_4 L_{nc}$$

GOALS

- **To reduce** the input to our language model
- **To better represent** many languages
- **To take advantage** of the multilingual setting

Hybrid Embedding Model (HEM)

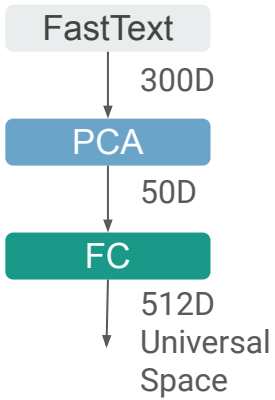


To reduce
the input to our language model

Hybrid Embedding Model (HEM)

Language Specific

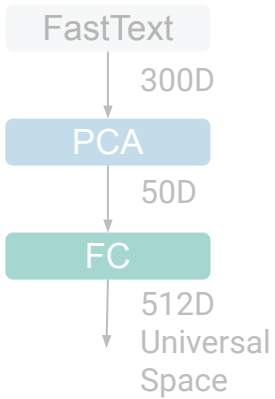
- Unique representations for common words
  - Top-5k most frequent words



Hybrid Embedding Model (HEM)

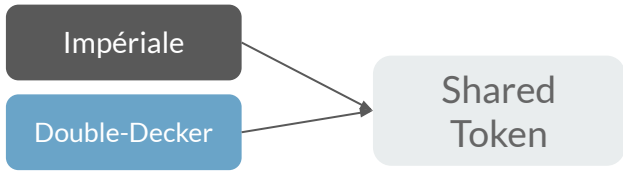
Language Specific

- Unique representations for common words
  - Top-5k most frequent words



Language Agnostic

- Shared vocab for uncommon words 40K tokens



Mapping is computed as a pretraining Step

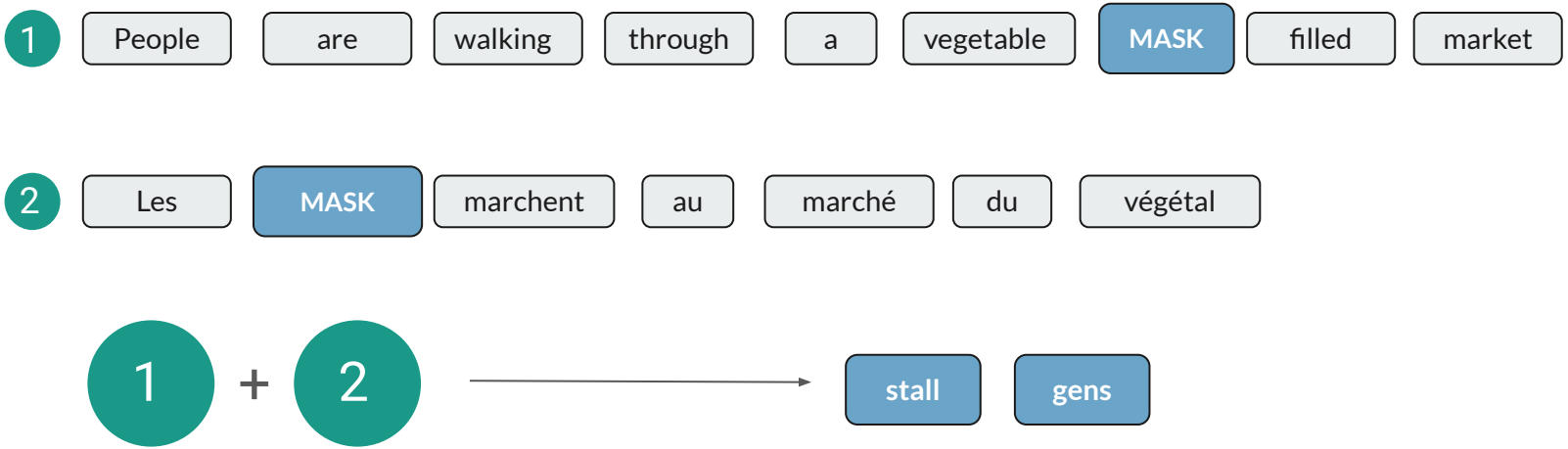
Explore When predicting latent token, randomly choose among top K scoring tokens *(instead of top 1)* with probability 0.2 (K = 20)

Masked Cross-Language Modeling (MCLM)

To better represent many languages

Masked Cross-Language Modeling (MCLM)

Randomly mask words in **2 semantically related sentences**



Cross Lingual Consistency (CLC)

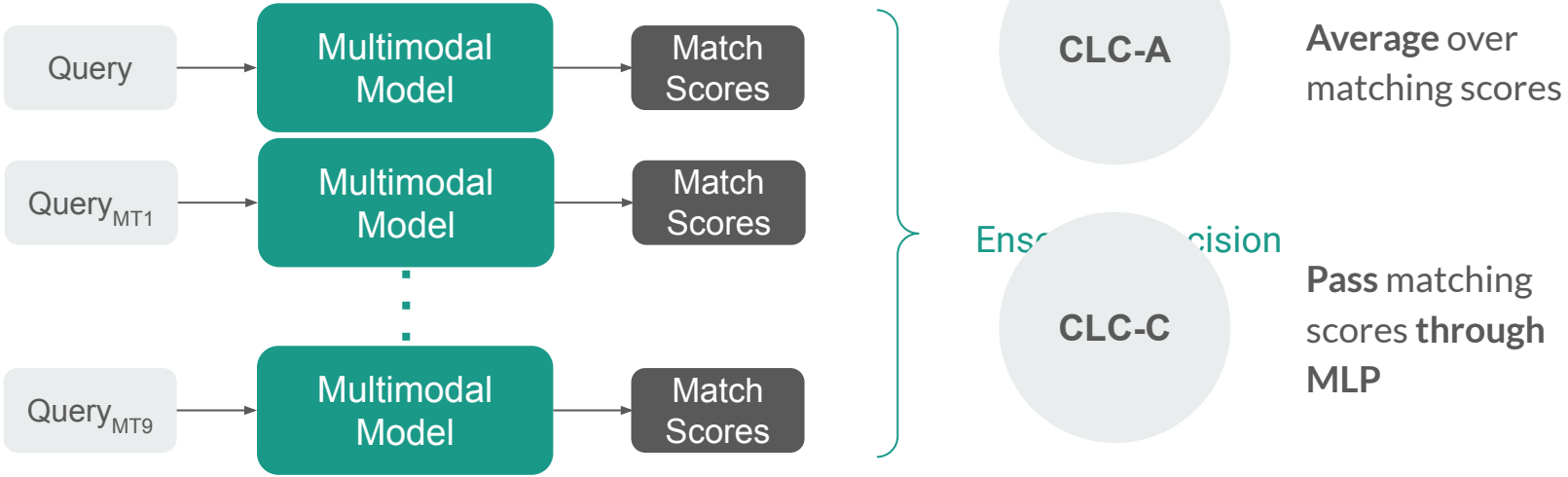
To take advantage of the multilingual setting

Cross Lingual Consistency (CLC)

Aggregate model decisions made from semantically similar sentences in different languages (sentence to image direction)

Cross Lingual Consistency (CLC)

Aggregate model decisions made from semantically similar sentences in different languages (sentence to image direction)



Cross Lingual Consistency (CLC)

Aggregate model decisions made from semantically similar sentences in different languages (sentence to image direction)

Cn: 很多 球迷 都 在 看 台上 观看 棒球 比赛

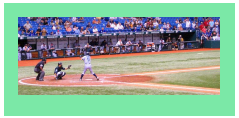
Cn → En: many fans in the stands to watch a baseball game

Cn → Fr: de nombreux fans dans les tribunes pour regarder un match de baseball

Cn → Ru: многие болельщики на трибунах, чтобы посмотреть игру в бейсбол

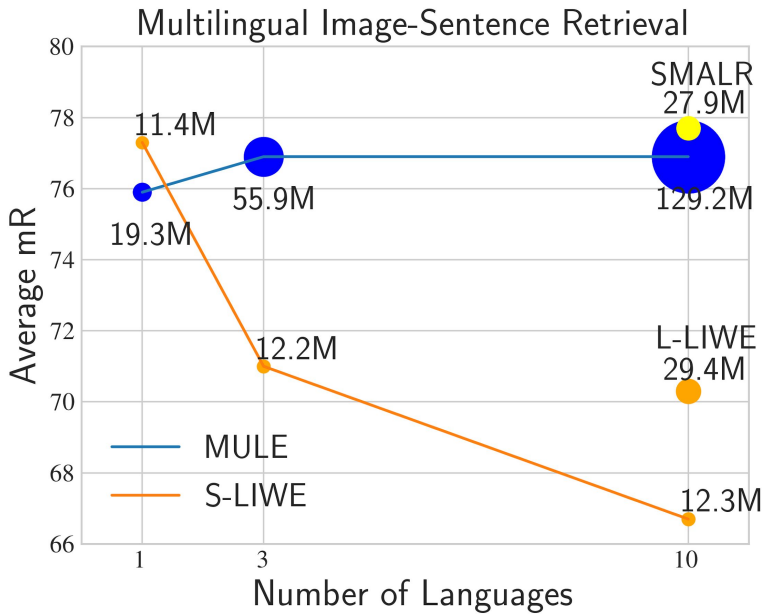
+ CLC-C

Cn: 很多 球迷 都 在 看 台上 观看 棒球 比赛



Final Thoughts

- **SMALR** uses **5X fewer parameters** than competitor MULE, and outperforms scalable LIWE
- Evaluated on **10 languages**
- **HEM** reduces input to language model
- **MCLM** aligns many diverse languages
- **CLC** better aggregates retrieved results, taking advantage of the multilingual setting



Vision-Language Introduction Summary

- Vision-language tasks have some key distinctions from vision-only problems
  - Many ways to describe visual information, no one way is ‘best’
  - Long tail distribution in language
  - Relating language and visual data with abstract, commonsense reasoning
- Common tasks, architectures, losses
  - Image-Sentence Retrieval, Triplet Loss
  - Image Captioning, Negative Log Likelihood
  - Visual Question Answering, Cross Entropy
- Recent directions of vision-language representation learning
  - Extending to multiple languages
  - Scalability vs. performance

Thank you!

Questions?