Multimodal Machine Translation

Lucia Specia

University of Sheffield

l.specia@sheffield.ac.uk

MTM - Lisbon, 1 Sept 2017

A wall divided the city.

A wall divided the city.

A wall divided the city.

$\rightarrow$ Eine **Wand** teilte die Stadt.

A wall divided the city.

A wall divided the city.

→ Eine **Mauer** teilte die Stadt.

Overview

Overview

Scope

Brazil opens vast Amazon reserve to mining

6 hours ago | Latin America & Caribbean | 135 Share

The reserve was created in 1984 by the then-military government

Brazil's government has abolished a vast national reserve in the Amazon to open up the area to mining.

The area, covering 46,000 sq km (17,800 sq miles), straddles the northern states of Amapa and Para, and is thought to be rich in gold, and other minerals.

The government said nine conservation and indigenous land areas within it would continue to be legally protected.

But activists have voiced concern that these areas could be badly compromised.

Scope

Brazil opens vast Amazon reserve to mining

6 hours ago | Latin America & Caribbean | 135 Share

The reserve was created in 1984 by the then-military government

Brazil's government has abolished a vast national reserve in the Amazon to open up the area to mining.

The area, covering 46,000 sq km (17,800 sq miles), straddles the northern states of Amapa and Para, and is thought to be rich in gold, and other minerals.

The government said nine conservation and indigenous land areas within it would continue to be legally protected.

But activists have voiced concern that these areas could be badly compromised.

Machine Translation

Text Summarisation

Text Simplification

(Natural Language Generation)

Hypothesis

Humans

Use a lot more cues than just text when making sense of the world and performing tasks

Hypothesis

Humans

Use a lot more cues than just text when making sense of the world and performing tasks

Image can contribute in cases of

Ambiguity (lexical, gender, syntactic)
Vagueness
OOV
Relevance, etc

Hypothesis

Humans

Use a lot more cues than just text when making sense of the world and performing tasks

Image can contribute in cases of

Ambiguity (lexical, gender, syntactic)
Vagueness
OOV
Relevance, etc

Vision & language very popular nowadays

Annual workshops since 2011
Tutorials since 2013
Summer schools since 2015, etc

Overview

Background

Work on **language grounding**:

- Images to represent a model of **perception** of the world:
  - Train a CNN on a object recognition task, e.g. [Xu et al., 2015]
  - Do a forward pass given an image input
  - Use one or more layers (e.g. FC_7, $CONV_5$) or output for language task

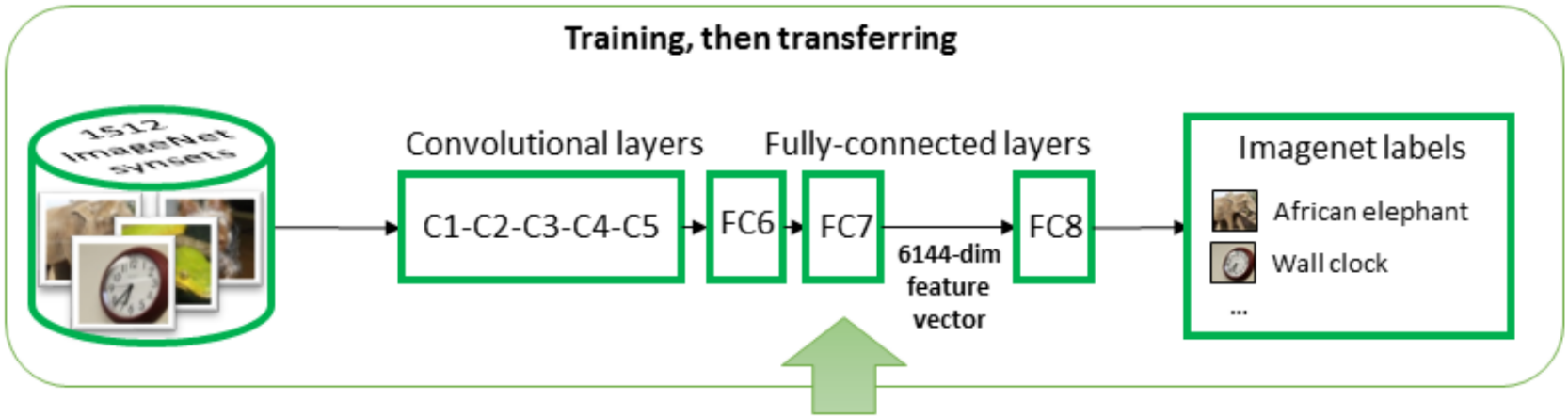
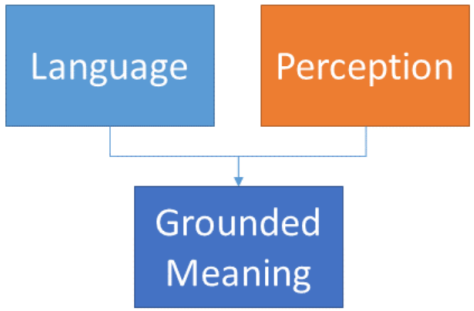


Image from (Elliott et al., ACL16) tutorial on Multimodal Learning and Reasoning

Background - Language grounding

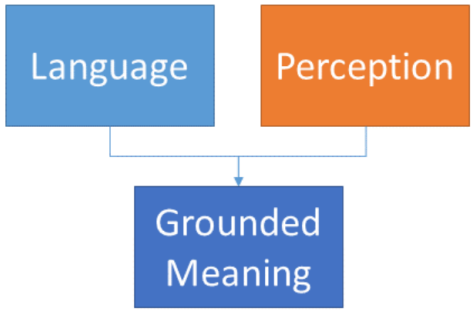
Representational grounded (lexical) semantics



- **Multimodal semantics** to represent the meaning of a word
- Method: Fusion

Background - Language grounding

Representational grounded (lexical) semantics



- **Multimodal semantics** to represent the meaning of a word
- Method: Fusion

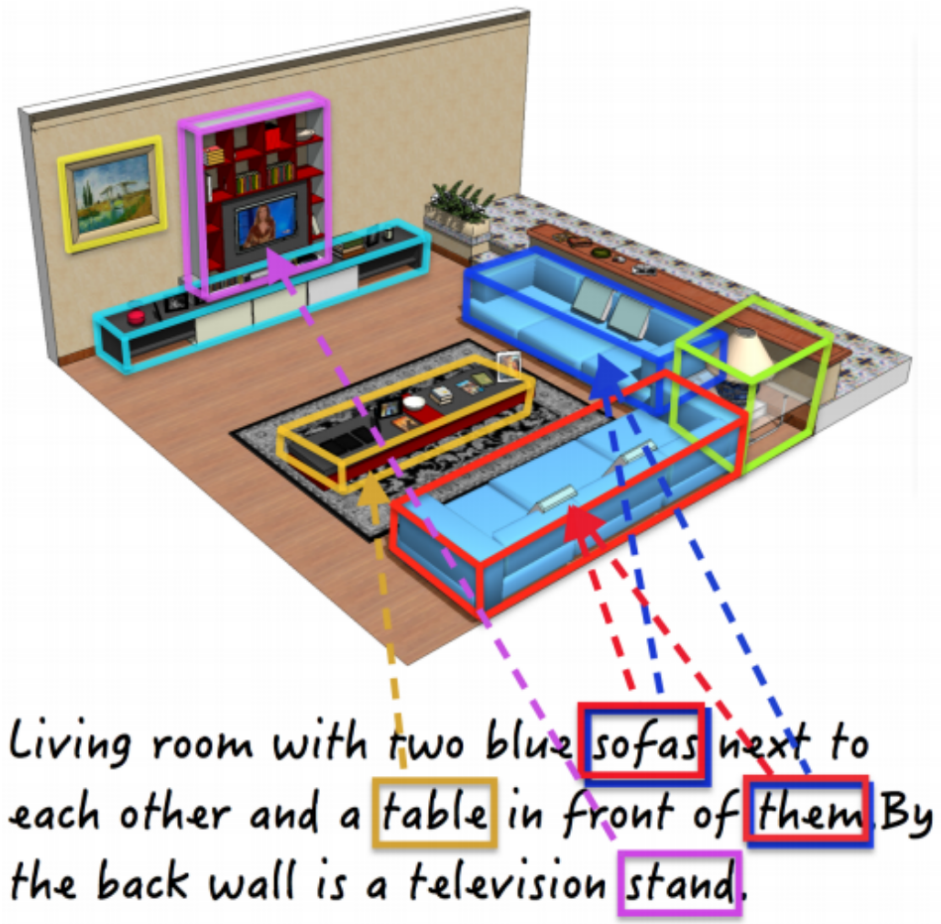
Referential grounded (lexical) semantics

- **Cross-modal semantics** to determine the referent a word denotes
- Method: Mapping

Images from (Elliott et al., ACL16) tutorial on Multimodal Learning and Reasoning

Background - Referential grounding

Idea of **mapping**:



Images from (Elliott et al., ACL16) tutorial on Multimodal Learning and Reasoning

Overview

Background

Monolingual work in **Computer Vision**:

Girl in pink dress is jumping in air.

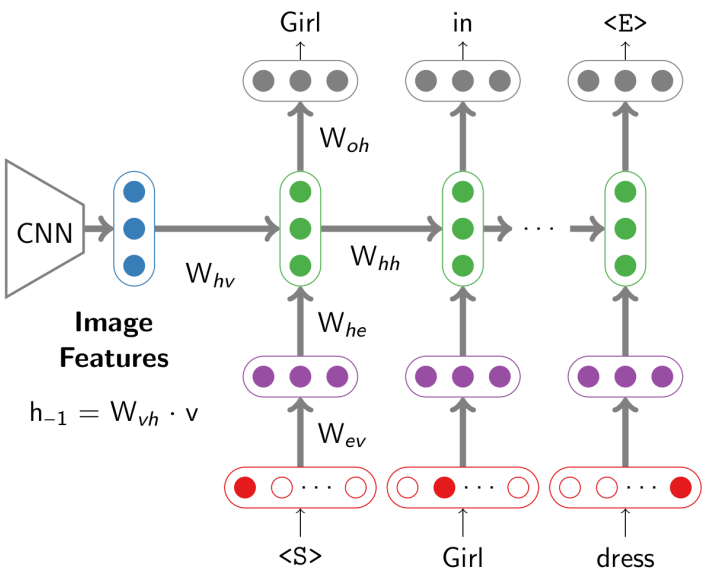
- Image captioning

Background

Monolingual work in **Computer Vision**:

Girl in pink dress is jumping in air.

- Image captioning



Prediction
$\text{softmax}(o_i)$

Output
$o_i = h_i \cdot W_{oh}$

Recurrent
$h_i = f(h_{i-1}, e_i)$

Embeddings
$e_i = x_i \cdot W_{ev}$

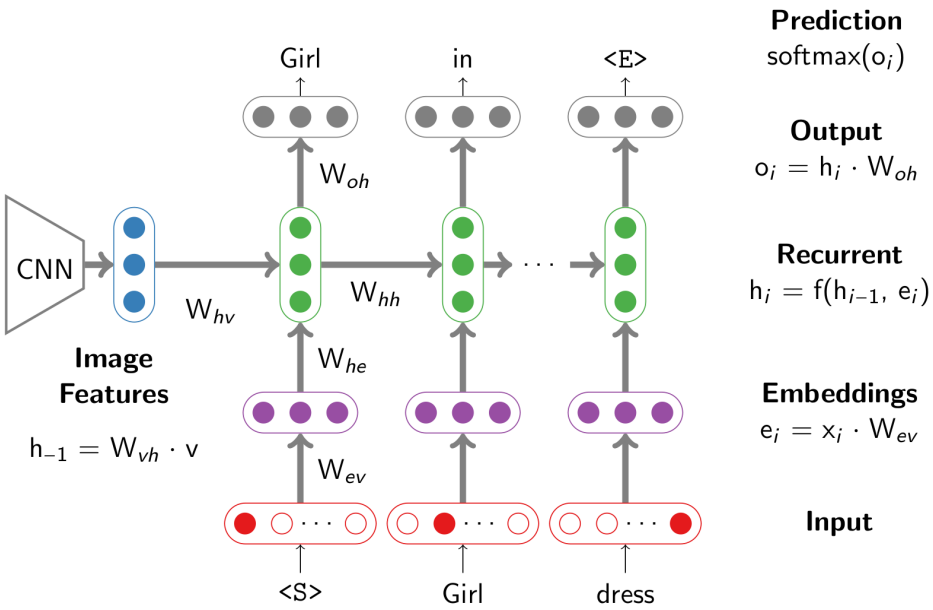
Input

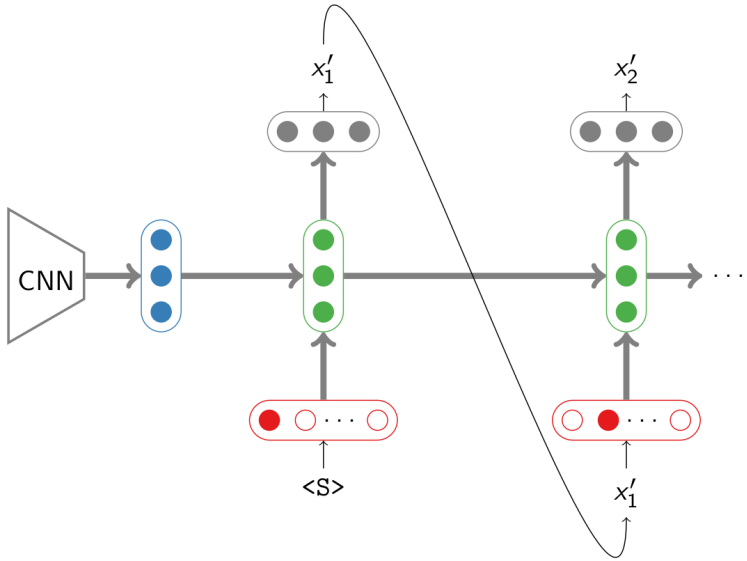
Background

Monolingual work in **Computer Vision**:

Girl in pink dress is jumping in air.

- Image captioning





Images from (Elliott et al., ACL16) tutorial on Multimodal Learning and Reasoning

Background - Computer Vision

- Visual question answering

Yellow

What colour is the moustache made of?

Background - Computer Vision

- Visual question answering

Yellow

What colour is the moustache made of?

Q: Describe what happened immediately after this picture was taken.

A: They drove around.

Background - Computer Vision

- Visual question answering

Yellow

What colour is the moustache made of?

Q: Describe what happened immediately after this picture was taken.

A: They drove around.

- Video captioning
- Scene description, etc.

Images from (Elliott et al., ACL16) tutorial on Multimodal Learning and Reasoning

Overview

Multimodal Machine Translation

Given a **text** which has one or more **images** associated with it:

Offshore ship gets 'seal of approval' from rescued pup

29 November 2011 | NE Scotland, Orkney & Shetland

The pup climbed back on board and fell asleep

A tired seal pup which refused to leave an oil industry ship 100 miles out to sea is recovering before being released back into the wild.

British Divers Marine Life Rescue (BDMLR) was alerted that the grey seal pup had climbed on board MSV Subsea Viking west of Shetland.

The crew placed the pup back in the water - but it returned, and fell asleep on board.

Initially named Sammy, Viking was taken to BDMLR's Highland Seal Hospital.

It is thought the pup would have been exhausted as they cannot swim very well while they still have their white coat.

Ali Jack, of BDMLR, said: "The poor wee thing was exhausted and if the ship's crew had not rescued this pup when they did then it would have succumbed to its

Find **alignments** (i.e. mappings):

Offshore ship gets 'seal of approval' from rescued pup

29 November 2011 | NE Scotland, Orkney & Shetland

The pup climbed back on board and fell asleep

A tired seal pup which refused to leave an oil industry ship 100 miles out to sea is recovering before being released back into the wild.

British Divers Marine Life Rescue (BDMLR) was alerted that the grey seal pup had climbed on board MSV Subsea Viking west of Shetland.

The crew placed the pup back in the water - but it returned, and fell asleep on board.

Initially named Sammy, Viking was taken to BDMLR's Highland Seal Hospital.

It is thought the pup would have been exhausted as they cannot swim very well while they still have their white coat.

Ali Jack, of BDMLR, said: "The poor wee thing was exhausted and if the ship's crew had not rescued this pup when they did then it would have succumbed to its

Multimodal Machine Translation

Use **grounded** language as part of a translation model:

O navio a céu aberto obtém o "selo de aprovação" do cachorrinho resgatado

29 de novembro de 2011 | NE Escócia, Orkney & Shetland

Um filhote de foca cansado que se recusou a deixar um navio da indústria do petróleo 100 milhas para o mar está se recuperando antes de ser liberado de volta à natureza.

British Divers Marine Life Rescue (BDMLR) foi alertado de que o filhote de cachorro cinza escalara a bordo do MSV Subsea Viking ao oeste de Shetland.

A tripulação colocou o cachorrinho de volta na água - mas retornou e adormeceu a bordo.

Inicialmente chamado Sammy, Viking foi levado para o Hospital Highland Seal da BDMLR.

Pensa-se que o cachorrinho teria sido exausto porque não pode nadar muito bem enquanto ainda tem o casaco branco.

Ali Jack, do BDMLR, disse: "A pobre coisa estava exausta e se a tripulação do navio não tivesse resgatado este filhote quando eles o fizeram, teria sucumbido ao seu cansaço e provavelmente se afogou.

"Gostaríamos de prestar homenagem especial à tripulação por ir ao comprimento que eles fizeram para salvar a vida deste filhote".

Challenges

1. **Object detection** is not perfect and strongly biased towards objects seen in training
2. **Mapping models** only work well enough in closed domains
3. No obvious way to **encode sparse image information** along with language models
4. No large enough **multimodal dataset** to train translation models

Challenges

1. **Object detection** is not perfect and strongly biased towards objects seen in training
2. **Mapping models** only work well enough in closed domains
3. No obvious way to **encode sparse image information** along with language models
4. No large enough **multimodal dataset** to train translation models

Solutions:

- Translate image description datasets
- Use dense, low-level intermediate layer CNN features

Challenges - Object detection

ImageNet

- Image database organised acc. to WordNet hierarchy (nouns)
- Synsets (or object "categories"): 21,841
- Number of images: 14,197,122 (average 500 per synset)
- Number of images with bounding box annotations: 1,034,908
- In practice, we use models trained on 1,000 object categories from **ILSVRC shared tasks** [Russakovsky et al., 2015]

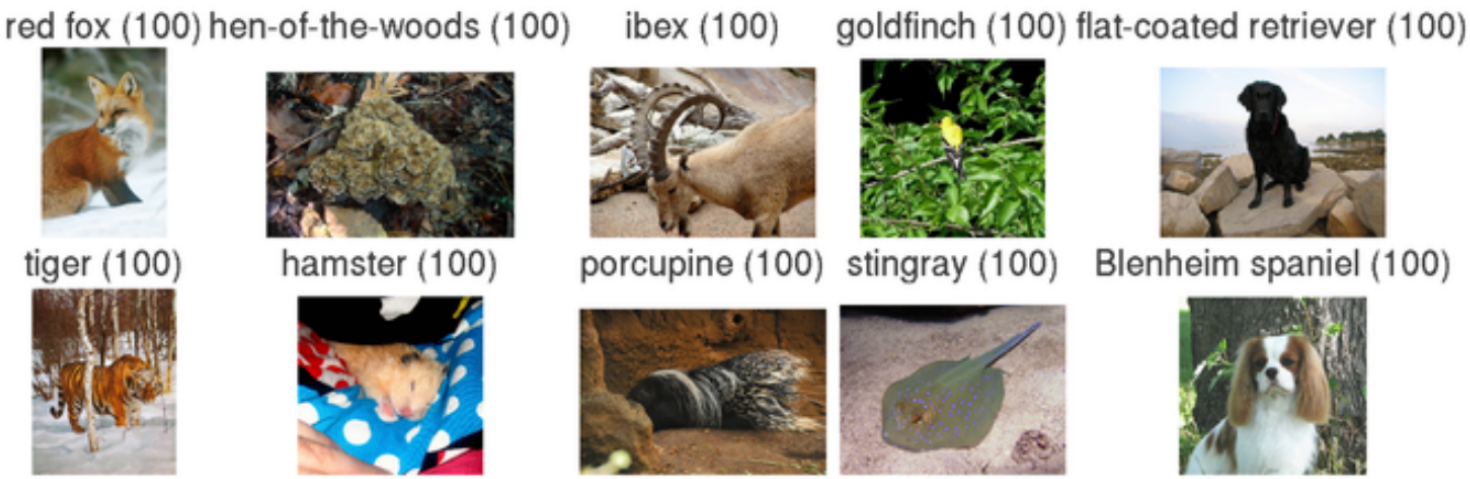
Challenges - Object detection

ImageNet

High level category	# synset (subcategories)	Avg # images per synset	Total # images
animal	3822	732	2799K
device	2385	675	1610K
person	2035	468	952K
plant	1666	600	999K
food	1495	670	1001K
structure	1239	763	946K
mammal	1138	821	934K
tree	993	568	564K
covering	946	819	774K
bird	856	949	812K
invertebrate	728	573	417K
fish	566	494	280K
vehicle	481	778	374K
flower	462	735	339K
tool	316	551	174K
fruit	309	607	188K
fungus	303	453	137K
reptile	268	707	190K
fabric	262	690	181K
furniture	187	1043	195K
vegetable	176	764	135K
sport	166	1207	200K
musical instrument	157	891	140K
geological formation	151	838	127K
amphibian	94	591	56K
utensil	86	912	78K
appliance	51	1164	59K

Challenges - Object detection

Top-10 easiest categories to predict [Russakovsky et al., 2014] from ImageNet (**ILSVRC**)



Challenges - Datasets

General texts make mapping too complex

- Use sentences that are **descriptions**: image captioning datasets
- Evidence that image description generation is "good enough"
- **Monolingual** datasets exist which can be extended to other languages

Challenge - Dataset creation

32.5K English→German/French images and professional translations from English Flickr30K [Elliott et al., 2016, Elliott et al., 2017]

Sentences and images

Training set	Development set	Test2016
29,000	1,014	1,000

Sentences and images

Test2017	TestCOCO
1,000	461

Challenge - Dataset creation

Flick30K

En: A group of people are eating noodles.

De: Eine Gruppe von Leuten isst Nudeln.

Fr: Un groupe de gens mangent des nouilles.

Challenge - Dataset creation

Ambiguous COCO (from Verse [Gella et al., 2016])

En: A red train is passing over a bridge

De: Ein roter Zug fährt auf einer Brücke über das Wasser

Fr: Un train rouge traverse l'eau sur un pont.

En: A man on a motorcycle is passing another vehicle

De: Ein Mann auf einem Motorrad fährt an einem anderen Fahrzeug vorbei.

Fr: Un homme sur une moto dépasse un autre véhicule.

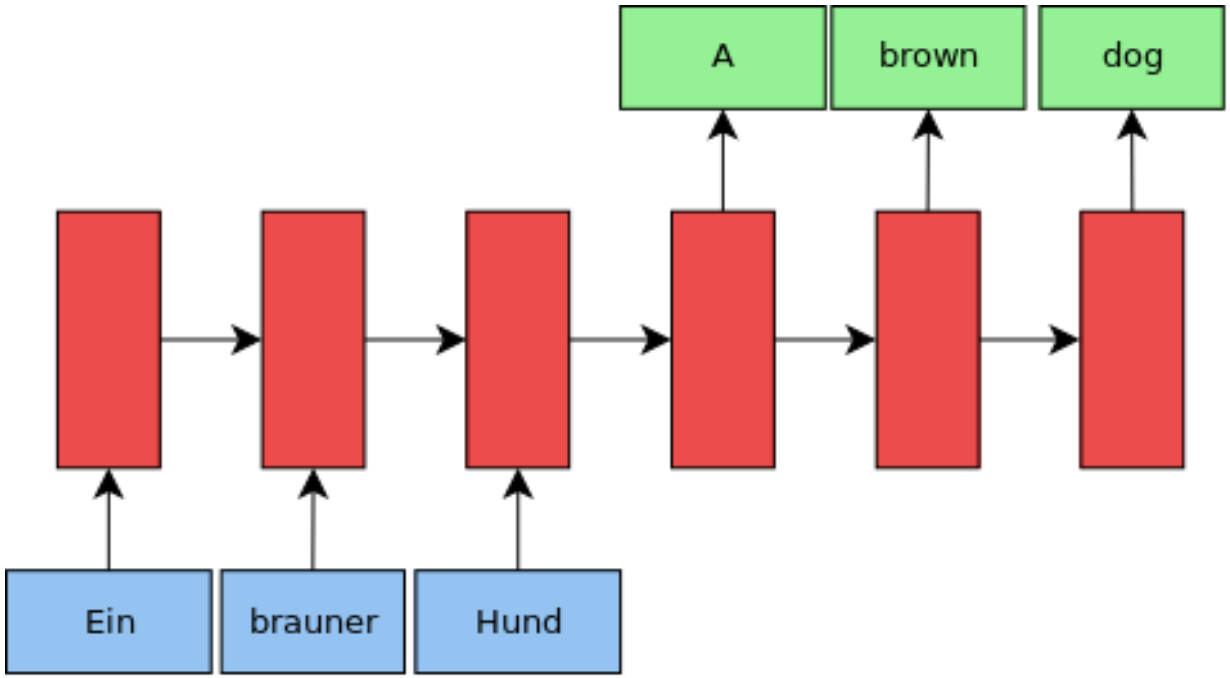
Overview

General framework

- Sequence-to-sequence (**encoder-decoder**) neural net models
- Visual information:
  - Dense, low-level feature vectors (layers of CNN)
  - **Less common**: sparse object categories (output of CNN)
- **Basic method**: visual information to initialise encoder/decoder/both, or concatenated with word representations (at each time step)

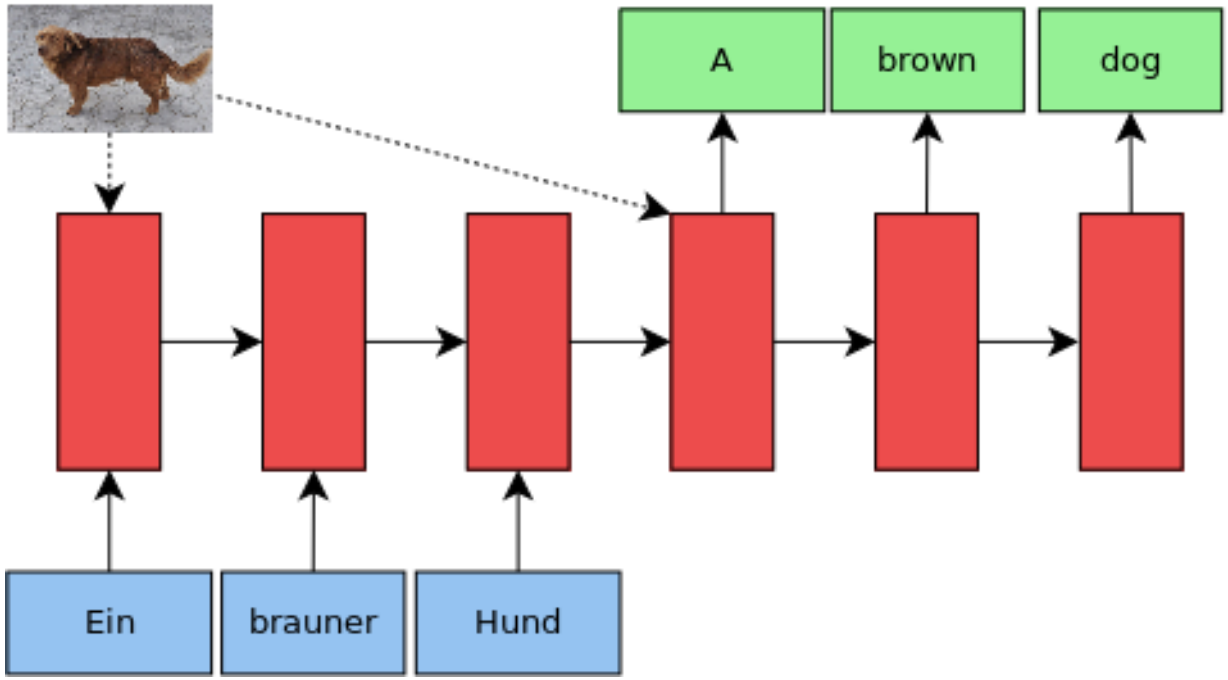
General framework

NMT $\rightarrow$ MMT



General framework

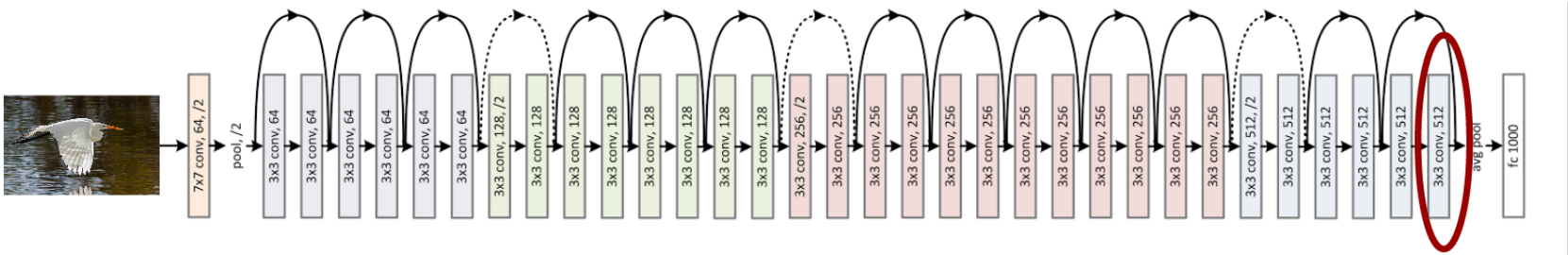
NMT $\rightarrow$ MMT



General framework

Visual info: intermediate layers from ResNet-50 CNN trained on ImageNet for object recognition task:

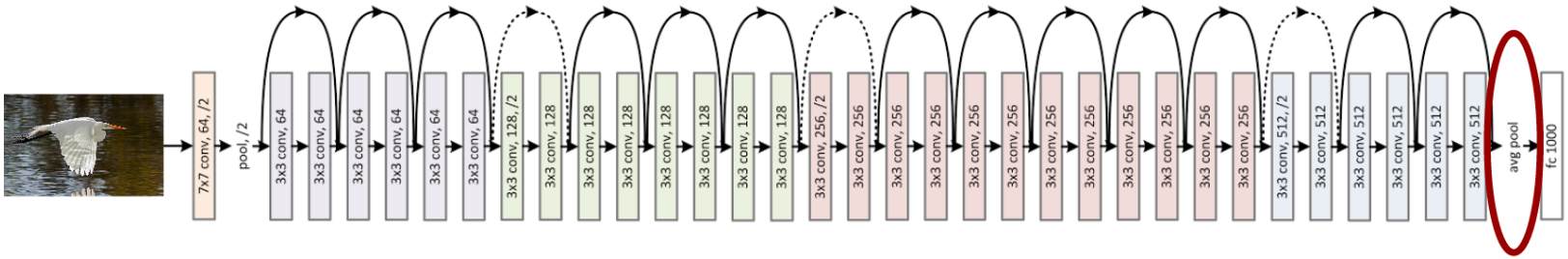
Last convolutional layer (14x14x1024D tensor)



General framework

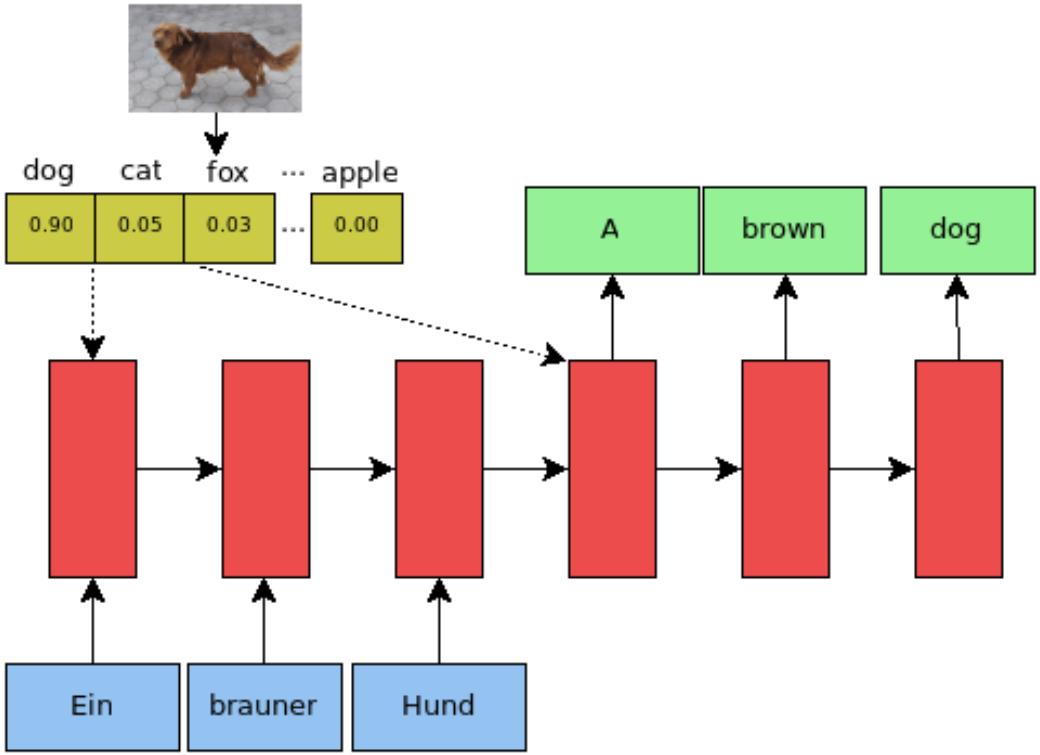
Visual info: intermediate layers from ResNet-50 CNN trained on ImageNet for object recognition task:

Pooled output of final convolutional layer (4096D vector)

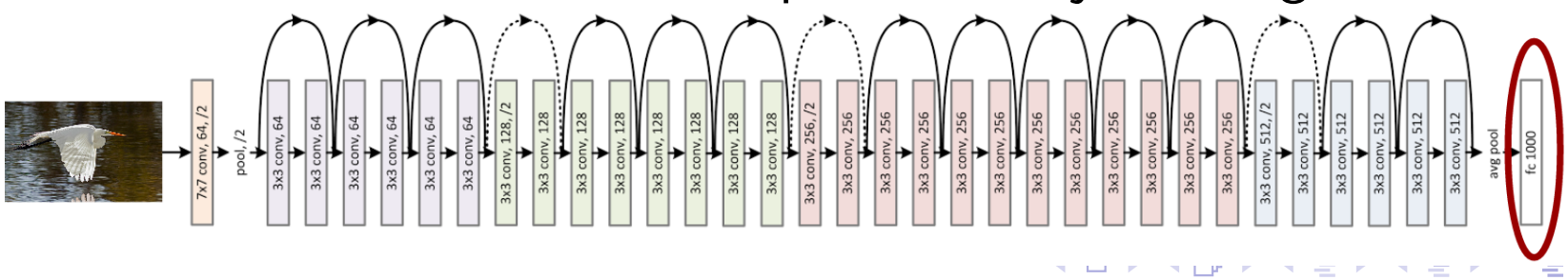


General framework

Alternative approach: "objects" rather than low-level features



Softmax: 1K vectors with predicted object categories



Overview

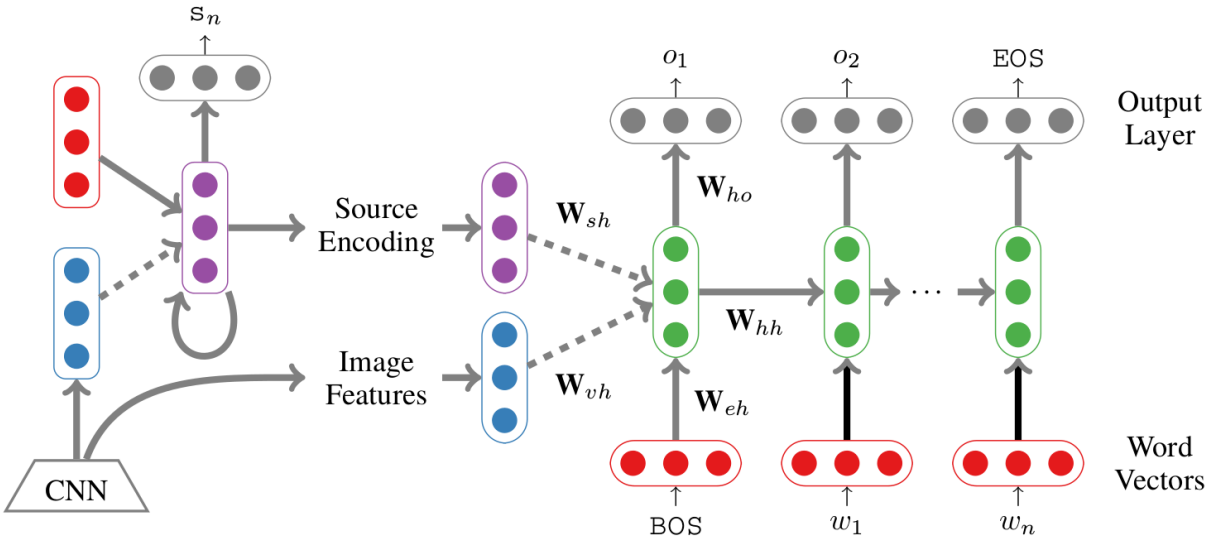
How well do systems do?

- WMT16 and WMT17 tasks on **Multimodal Machine Translation**[1]
- **Evaluation**:
  - **Meteor** against 1 reference translation
  - **Human evaluation** (direct assessments) (WMT17)

[1] http://www.statmt.org/wmt17/

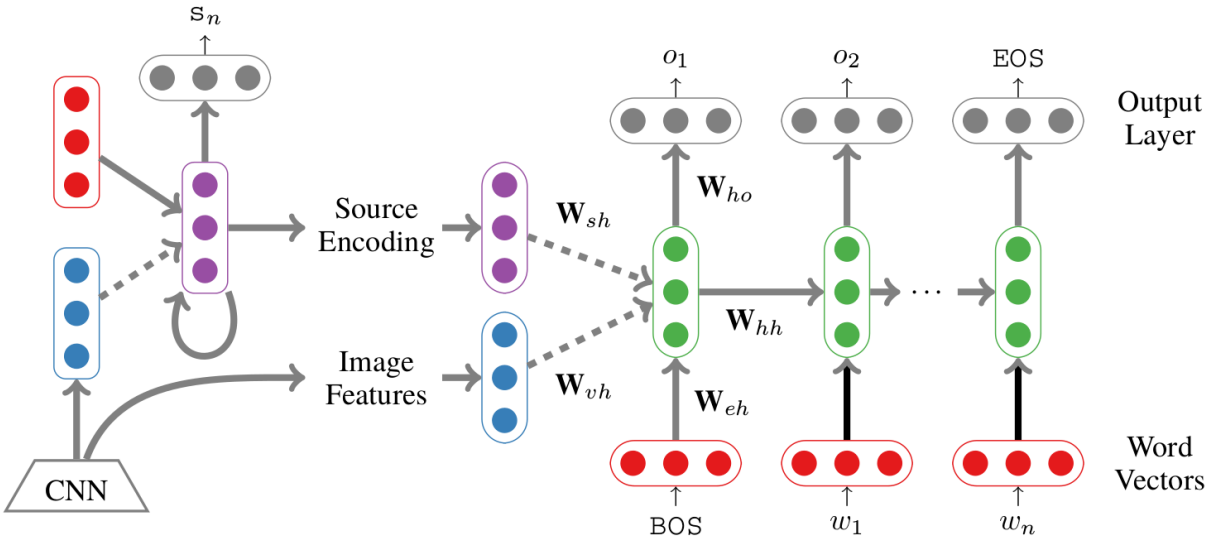
Results WMT16

- 10 teams submitted 16 systems
- Baselines:
  - **Grounded**: [Elliott et al., 2015] – the MLM→MLM



Results WMT16

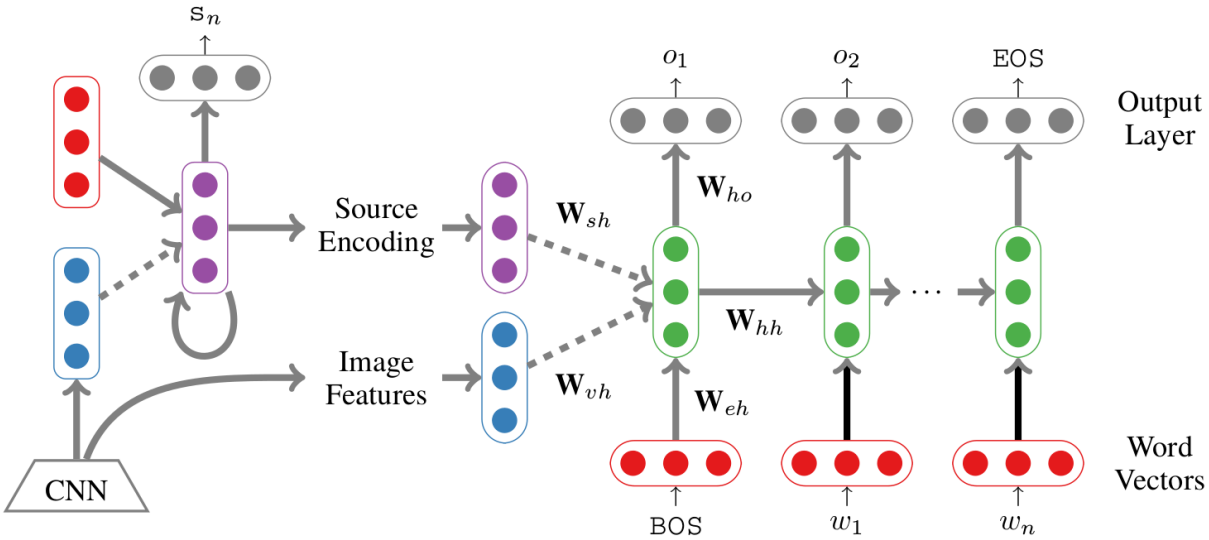
- 10 teams submitted 16 systems
- Baselines:
  - **Grounded**: [Elliott et al., 2015] – the MLM→MLM



  - **Moses**: PBSMT system trained on task parallel texts only

Results WMT16

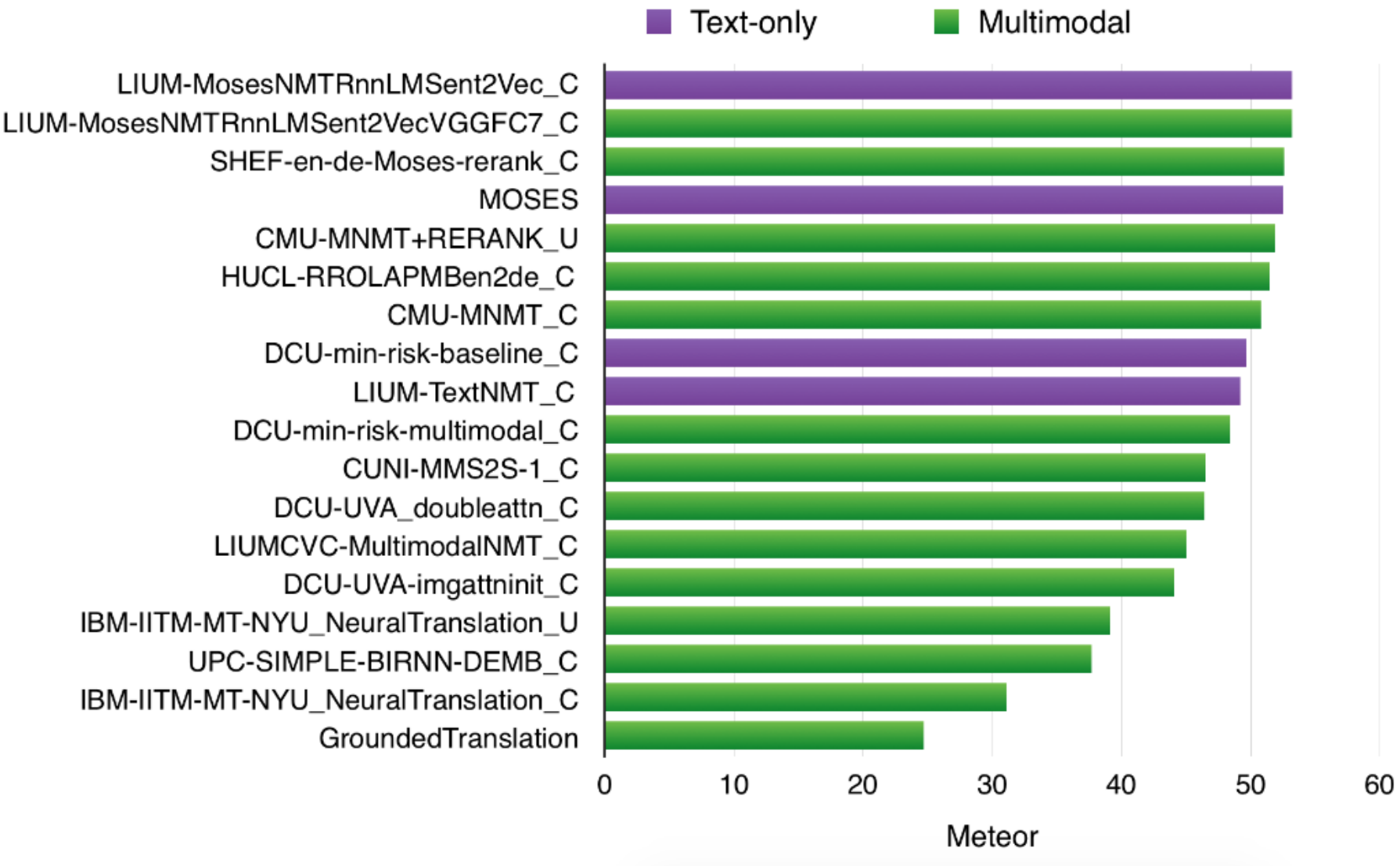
- 10 teams submitted 16 systems
- Baselines:
  - **Grounded**: [Elliott et al., 2015] – the MLM→MLM



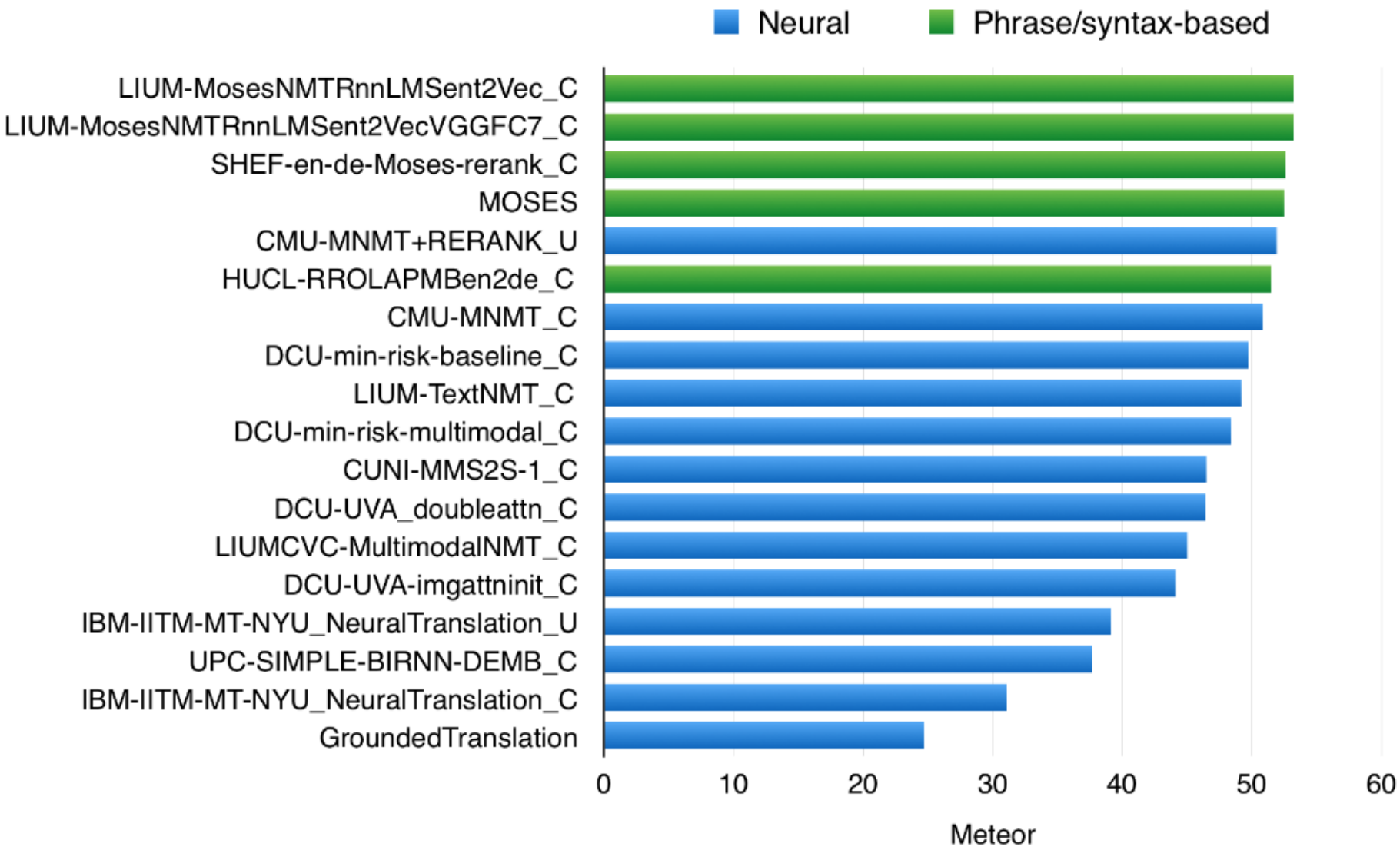
  - **Moses**: PBSMT system trained on task parallel texts only

- **Results**: multimodal systems did not do significantly better than monomodal ones

Results WMT16



Results WMT16



Results WMT16

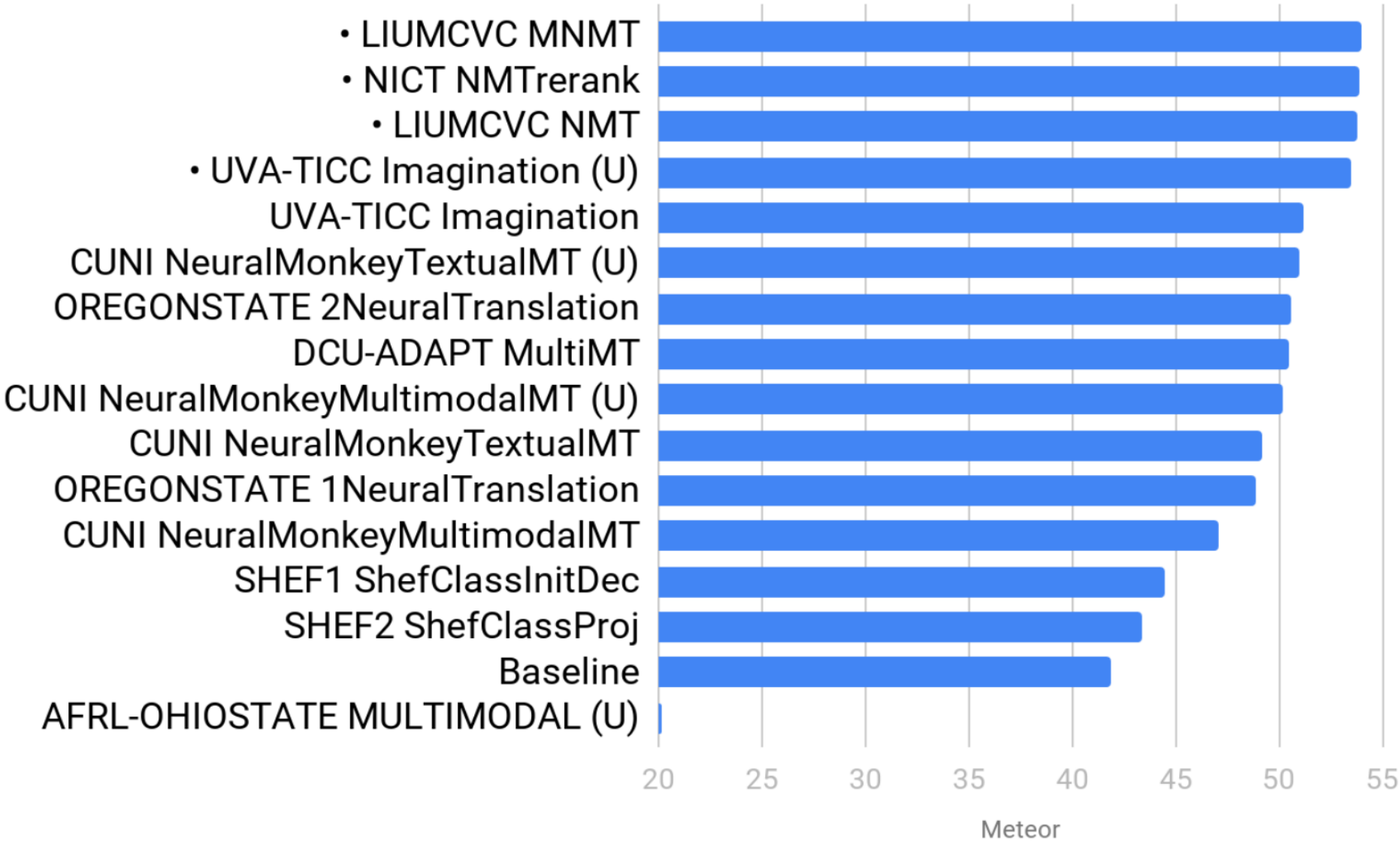
Multimodal integration techniques used:

- **NMT**: (attention-based) encoder-decoder approach
  - **CUNI** - input for decoder is linear combination of image features and two RNN encoders coding the source sentence and its translation from an SMT system
  - **CMU** - image features are appended in the head/tail of the textual features or dissipated in parallel LSTM threads.
  - **DCU** - image features used to initialise the target-side decoder
  - **IBM-IITM-Montreal-NYU** - same as Baseline except that image and source sentence representation are fed at every timestep to target RNN decoder
  - **LIUMCVC**: attention mechanism is shared across the two modalities

Results WMT17

- 9 teams submitted 15 systems
- Baseline:
  - Text-only **neural MT approach** (Nematus)
- **Results**: Some multimodal systems better than monomodal counterparts

Results WMT17 - Flickr17 (English-German) - Meteor



Results WMT17 - Flickr17 (English-German) - Human

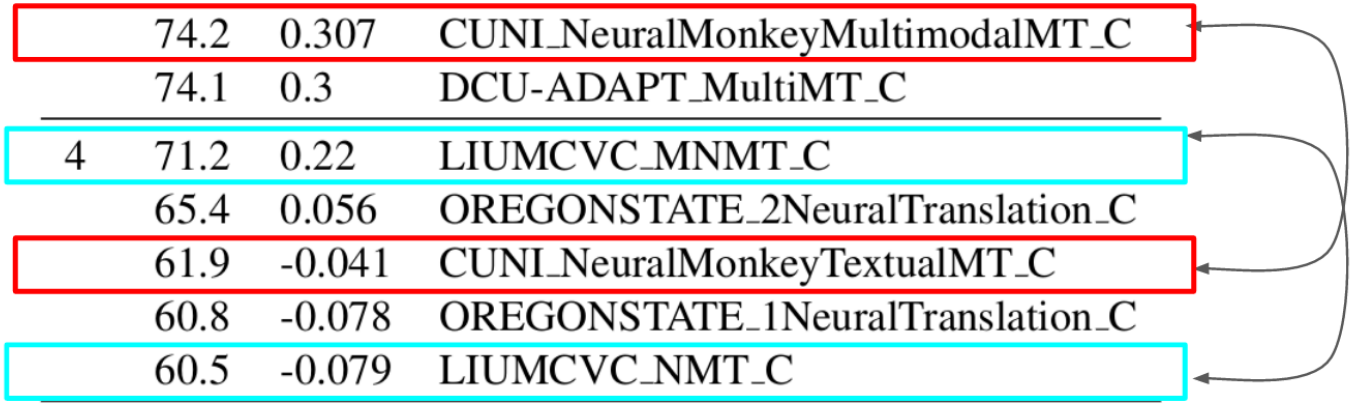
External resources helped

Visual context helped

#	Raw	z	System
1	77.8	0.665	LIUMCVC_MNMT_C
2	74.1	0.552	UvA-TiCC_IMAGINATION_U
3	70.3	0.437	NICT_NMTrerank_C
	68.1	0.325	CUNI_NeuralMonkeyTextualMT_U
	68.1	0.311	DCU-ADAPT_MultiMT_C
	65.1	0.196	LIUMCVC_NMT_C
	60.6	0.136	CUNI_NeuralMonkeyMultimodalMT_U
	59.7	0.08	UvA-TiCC_IMAGINATION_C
	55.9	-0.049	CUNI_NeuralMonkeyMultimodalMT_C
	54.4	-0.091	OREGONSTATE_2NeuralTranslation_C
	54.2	-0.108	CUNI_NeuralMonkeyTextualMT_C
	53.3	-0.144	OREGONSTATE_1NeuralTranslation_C
	49.4	-0.266	SHEF_ShefClassProj_C
	46.6	-0.37	SHEF_ShefClassInitDec_C
15	39.0	-0.615	Baseline (text-only NMT)
	36.6	-0.674	AFRL-OHIOSTATE_MULTIMODAL_U

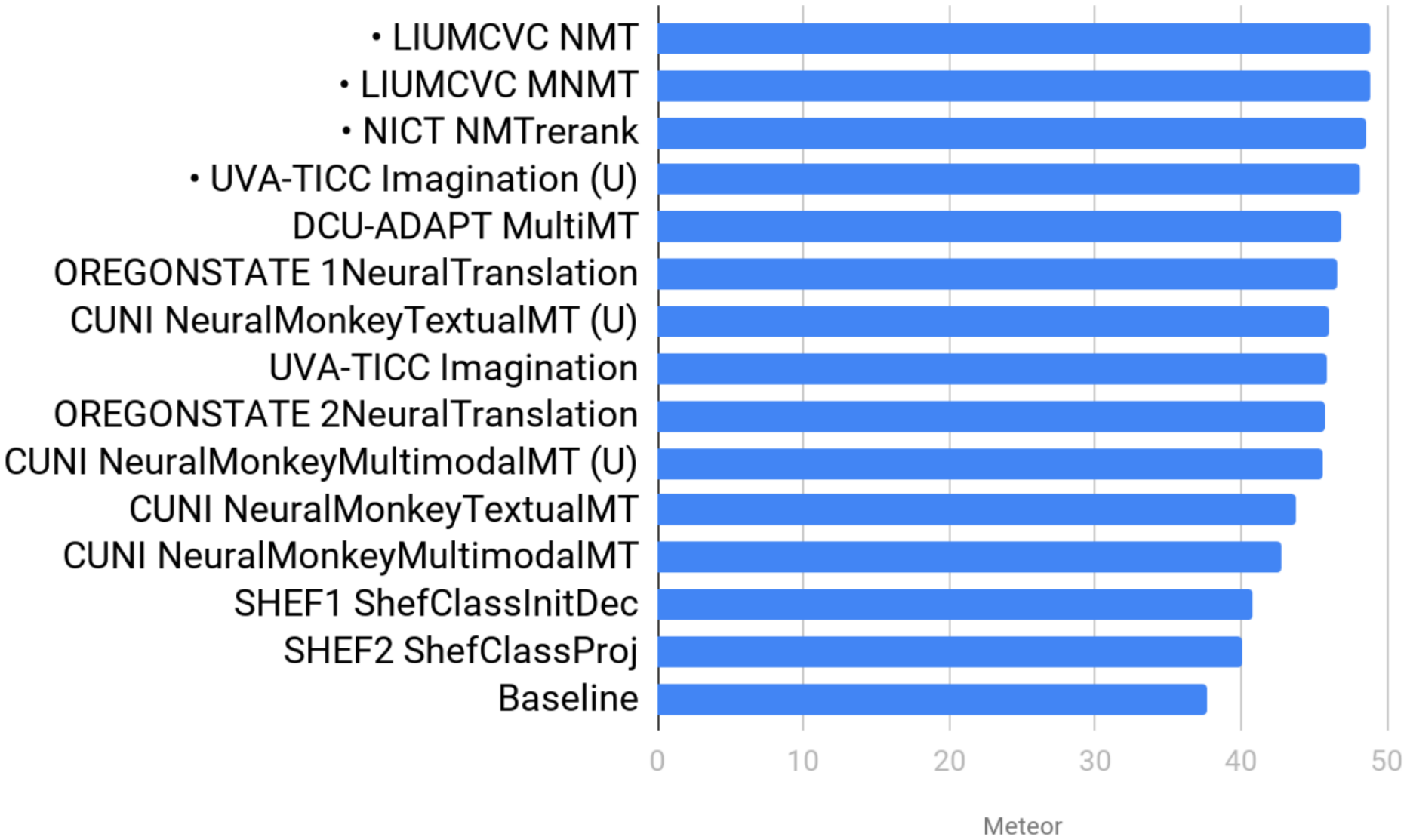
Results WMT17 - Flickr17 (English-French) - Human

			English→French
#	Raw	z	System
1	79.4	0.446	NICT_NMTrerank_C
	74.2	0.307	CUNI_NeuralMonkeyMultimodalMT_C
	74.1	0.3	DCU-ADAPT_MultiMT_C
4	71.2	0.22	LIUMCVC_MNMT_C
	65.4	0.056	OREGONSTATE_2NeuralTranslation_C
	61.9	-0.041	CUNI_NeuralMonkeyTextualMT_C
	60.8	-0.078	OREGONSTATE_1NeuralTranslation_C
	60.5	-0.079	LIUMCVC_NMT_C
9	54.7	-0.254	SHEF_ShefClassInitDec_C
	54.0	-0.282	SHEF_ShefClassProj_C
11	44.1	-0.539	Baseline (text-only NMT)

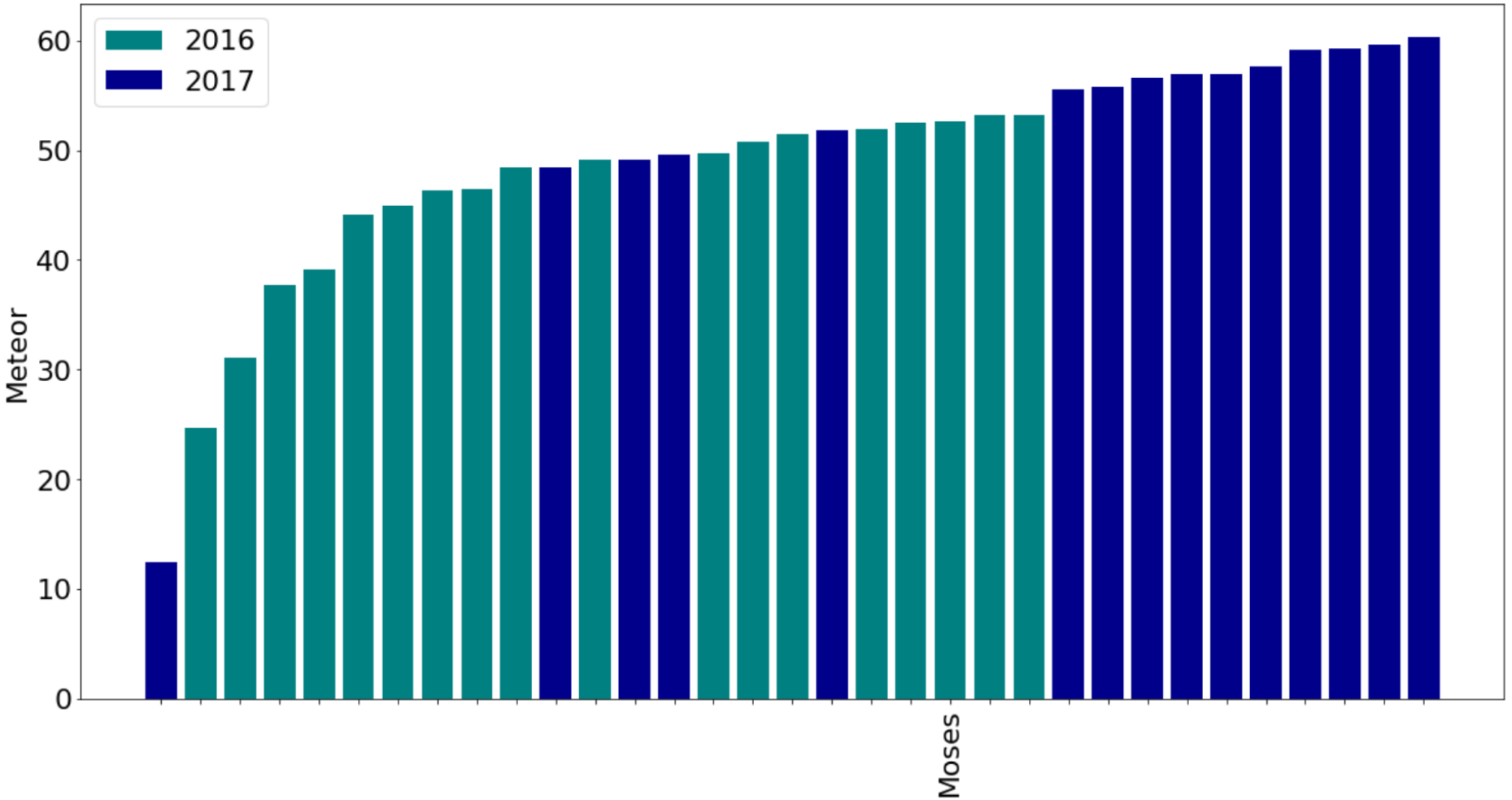


Visual context helped

Results WMT17 - Ambiguous COCO (English-German)



WMT17 vs WMT16 Systems - Flickr16 (English-German)



Summary of WMT17 systems

- Acc. to human evaluation, highest ranked systems are multimodal
- 3 types of submissions:
  - **Double-attention**: calculate context vectors over the source language hidden states and location-preserving feature vectors over the image; these vectors are used as input to decoder
  - **Encoder and/or decoder initialisation**: initialise the RNN with an affine transformation of a global image feature vector or initialising the encoder and decoder with the 1000 dimension softmax probability vector over the object classes
  - **Alternative**: element-wise multiplication of the target language embeddings with an affine transformation of a global image feature vector; summing the source language word embeddings with affine-transformed 1000 dimension softmax probability vector; using the visual features in a retrieval framework; and **learning visually-grounded encoder representations by learning to predict the global image feature vector from the source language hidden states**

Overview

How can we do better?

Still relatively small improvements over text-only NMT systems

- Better, more interesting data
- Better understanding of representations and their utility

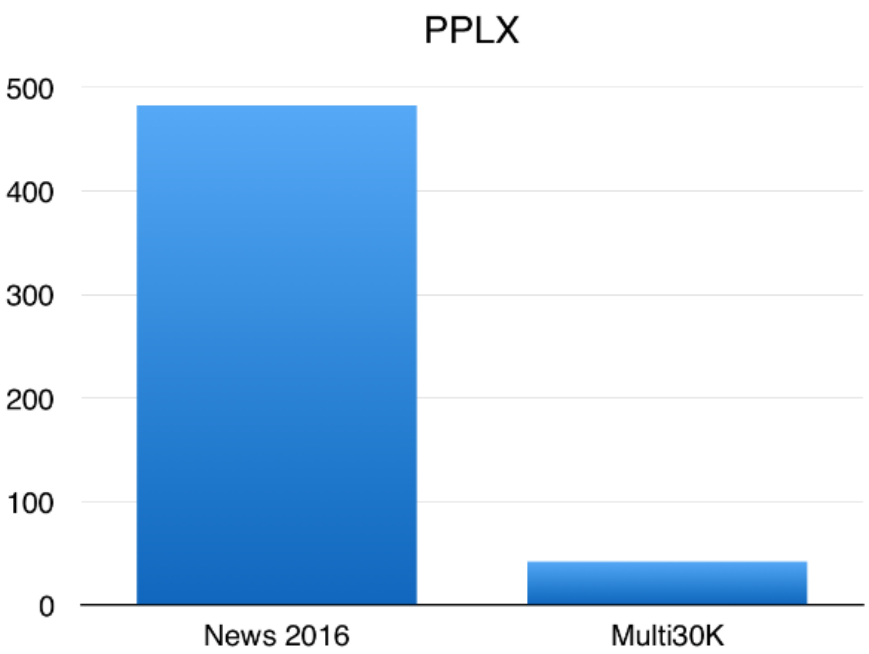
Better data

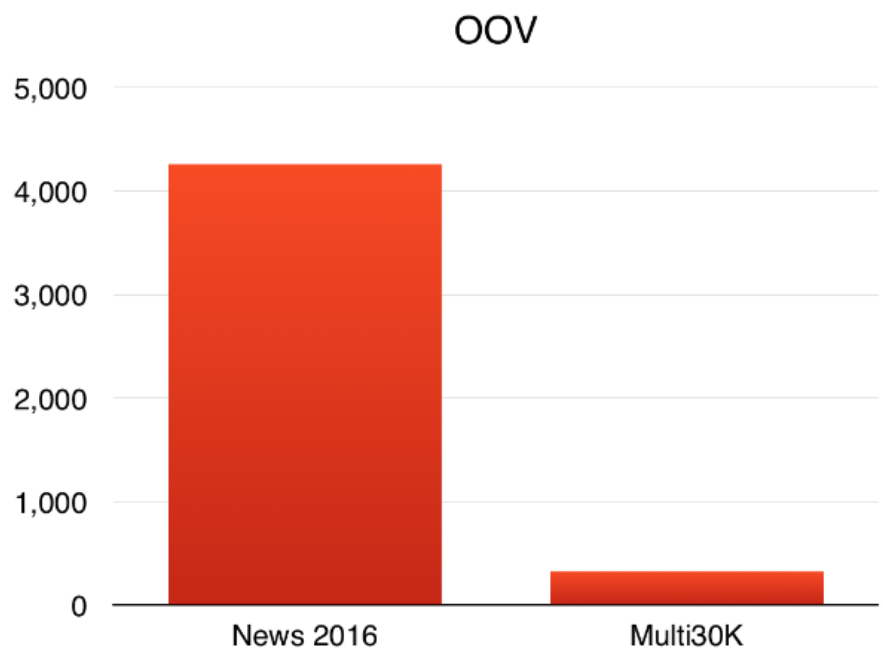
- Data is very simple: **text-only MT good enough**

		Sents.	Tokens	Types	TTR	Avg. length
WMT16	English	1,000	22,638	4,840	0.21	22.64
	German		22,429	5806	0.26	22.43
Flickr16	English	1,000	12,968	1,898	0.15	12.97
	German		12,103	2,125	0.18	12.10
	French		13,988	1,987	0.14	13.99
Flickr17	English	1,000	11,376	1,709	0.15	11.38
	German		10,758	2,153	0.20	10.76
	French		12,596	1,880	0.15	12.60
COCO17	English	461	5,239	953	0.18	11.36
	German		5,158	1,152	0.22	11.19
	French		5,710	1,104	0.19	12.39

Better data

- Data is very simple: **text-only MT good enough**





Better data

A woman is standing beside a bicycle with a dog.

Better data

How **Flickr/MSCOCO** and other IC datasets were collected:

- For specific object categories or human activities/scenes (e.g. **8 Flickr groups** in Flickr30K)
- Written by MTurk users in a **constrained setting**: "Describe what is happening in the picture" without using more than *n* characters
  - Repetitive vocabulary
  - Short sentences
  - Simple grammar
- Do not occur in the real world, less meaningful in real-life applications
- **Too small**: biggest has 80K descriptions
- To make them **multilingual**: translations are costly

Better data

A new multimodal, multilingual dataset[2]

- Rather than starting from images (+ contrived descriptions) and then translating them...
- Start from already existing multilingual corpora and find images to go with them
- **Movie subtitles**: Translations already available in multiple languages, find images to illustrate segments of the text

Mas quando se compra um carro vermelho, com estofamento preto...

But when you buy a red car with a black interior ...

✓ similar ✓ similar ◐ partially similar ✓ similar ✓ similar ✗ not similar

[2] Joint work with Josiah Wang

Better data

Children's Movies (SubIMDB)

But when you buy a red car with a black interior ...

OpenSubtitles

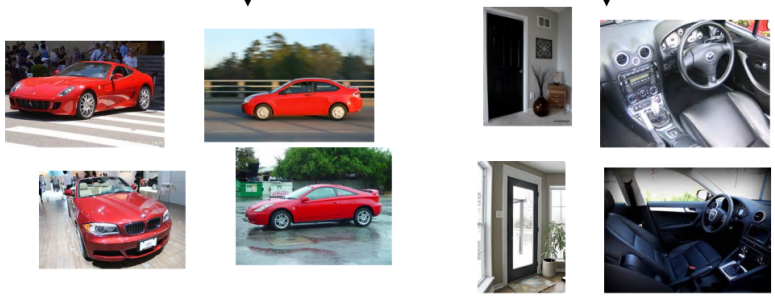
Mas quando se compra um carro vermelho, com estofamento preto...

Phrase extraction

a red car

a black interior

Image collection



Annotation

But when you buy a red car with a black interior ...

Better data

Queries

- Noun phrases (for now)
  - a white and beautiful bird ✓
  - pet bird ✓
  - a bird ✗
- 250,496 NP instances from 241,391 subtitle snippets
- 134,454 unique NPs
- Tokens per NP: 2–12 (mean 2.73)
- NPs ranked by concreteness score

Better data

Images

- Query unique NP using Bing Image Search
- "Free to share and use" CC license
- Download thumbnail images for top 150 results per query
- 131,085 NPs has at least one image (3,369 none)

Better data

Human evaluation

- Evaluate effectiveness of subtitle-image pair collection method
- Can be used as dataset for training classifiers to automatically annotate more data
- "Correct" instances can be used as dataset in (multilingual) multimodal tasks

Human evaluation

You have annotated 170 pieces of text so far!

Pick all images (if any) suitable for illustrating *at least one fragment* of the following piece of text:
(Detailed guidelines)

covered the hundred-acre wood .





I don't understand or
The text is too abstract to be illustrated or
I cannot think of **any** possible way that this piece of text can **ever** be illustrated by **any** image.

Comments (Optional):

Save & Next!

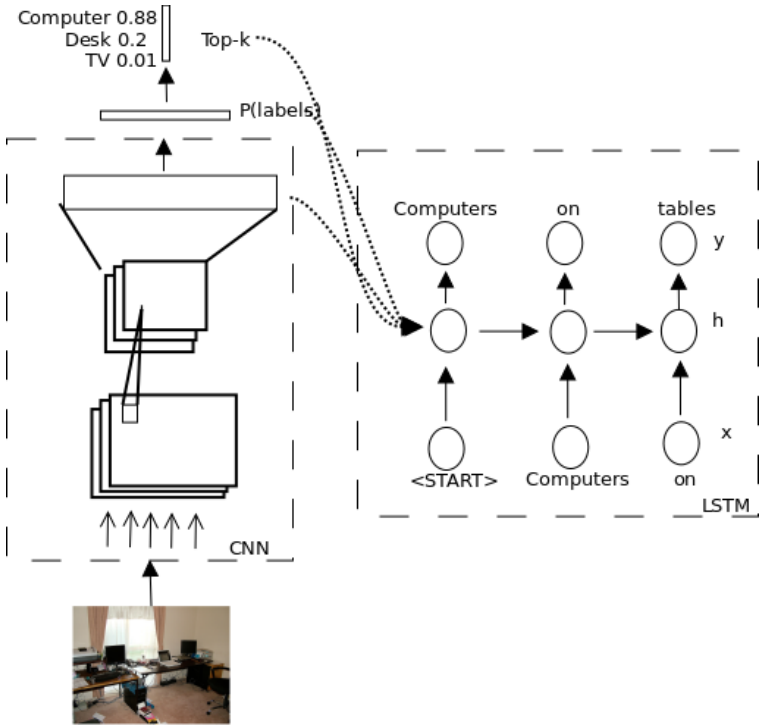
Better data

Human evaluation

- Subsampled 10,000 instances for annotation
  - Containing concrete NPs, high diversity, longer subtitles
  - Top 5 images each
- Annotate as **similar**, **partially similar**, **not similar**
- 10,000 subtitle snippet x 5 images each = 50,000 instances
- 646 unique annotators, all doubly annotated (2000 triply)
- **Results**: 32.8% similar, 9.8% partially similar, 57.4% not similar
- Ongoing work: improving phrase selection & filtering retrieved images

Understanding representations - MMT

How different representations contribute to MMT?[3]



- PENULTIMATE layer from a pre-trained CNN image network
- Class prediction vector (SOFTMAX): probability distribution over 1,000 object categories (ImageNet)

[3] Joint work with Pranava Madhyastha and Josiah Wang

Understanding representations - MMT

Image info to:

1. Initialising the encoder (**InitEnc**): images as the first token
2. Initialising decoder (**InitDec**): initialise the decoder's first hidden state with the predicted class distribution

Understanding representations - MMT

Performance on MMT task (WMT17 Flickr test sets):

Flickr	Feature	Model	Meteor	BLEU
EN–DE	–	Baseline	43.7	24.4
	PENULTIMATE	InitEnc	43.0	23.5
		InitDec	44.3	24.6
	SOFTMAX	InitEnc	42.4	23.3
		InitDec	44.5	25.0
EN–FR	–	Baseline	62.2	44.2
	PENULTIMATE	InitEnc	61.1	43.5
		InitDec	61.0	43.4
	SOFTMAX	InitEnc	61.0	43.3
		InitDec	62.8	45.0

Understanding representations - MMT

Performance on MMT task (WMT17 Flickr test sets):

Flickr	Feature	Model	Meteor	BLEU
EN–DE	–	Baseline	43.7	24.4
	PENULTIMATE	InitEnc	43.0	23.5
		InitDec	44.3	24.6
	SOFTMAX	InitEnc	42.4	23.3
		InitDec	44.5	25.0
EN–FR	–	Baseline	62.2	44.2
	PENULTIMATE	InitEnc	61.1	43.5
		InitDec	61.0	43.4
	SOFTMAX	InitEnc	61.0	43.3
		InitDec	62.8	45.0

Conclusions:

- Highly sparse but abstract semantic image information seems more equally or more useful for MMT tasks (and image captioning)

Understanding representations - MMT

Performance on MMT task (WMT17 Flickr test sets):

Flickr	Feature	Model	Meteor	BLEU
EN–DE	–	Baseline	43.7	24.4
	PENULTIMATE	InitEnc	43.0	23.5
		InitDec	44.3	24.6
	SOFTMAX	InitEnc	42.4	23.3
		InitDec	44.5	25.0
EN–FR	–	Baseline	62.2	44.2
	PENULTIMATE	InitEnc	61.1	43.5
		InitDec	61.0	43.4
	SOFTMAX	InitEnc	61.0	43.3
		InitDec	62.8	45.0

Conclusions:

- Highly sparse but abstract semantic image information seems more equally or more useful for MMT tasks (and image captioning) – despite ImageNet bias

Overview

Humans and MT systems

- **SRC**: A woman wearing a **hat** is making bread.
- **TXT**: Eine Frau mit einer **Mütze** macht Brot.
- **IMG**: Eine Frau mit einem **Hut** macht Brot.

Humans and MT systems

- **SRC**: Three children in **football uniforms** of two different teams are playing **football** on a **football field**, while another player and an adult stand in the background.
- **TXT**: Drei Kinder in **Fußballtrikots** zweier verschiedener Mannschaften spielen **Fußball** auf einem **Fußballplatz** während ein weiterer Spieler und eine Erwachsener im Hintergrund stehen.
- **IMG**: Drei Kinder in **Footballtrikots** zweier verschiedener Mannschaften spielen **Football** auf einem **Footballplatz** während ein weiterer Spieler und ein Erwachsener im Hintergrund stehen.

Humans and MT systems

- **MT**: Drei Kinder in Trikots spielen **Fußball** auf einem **Fußballfeld**, während ein anderer Spieler im Hintergrund stehen.
- **MMT**: Drei Kinder in Trikots spielen **Fußball** auf einem **Footballfeld**, während ein anderer Spieler und ein Erwachsener im Hintergrund spielen.

Humans and MT systems

- **SRC**: **A baseball player** in a black shirt just tagged a player in a white shirt.
- **TXT**: **Ein Baseballspieler** in einem schwarzen Shirt fängt einen Spieler in einem weißen Shirt.
- **IMG**: **Eine Baseballspielerin** in einem schwarzen Shirt fängt eine Spielerin in einem weißen Shirt.

Humans and MT systems

- **MT**: **Ein Baseballspieler** in einem schwarzen Hemd hat gerade einen Spieler in einem weißen Hemd.
- **MMT**: **Ein Baseballspieler** in einem schwarzen Hemd wirft ein Spieler in einem weißen Hemd.

Humans and MT systems

- **SRC**: One man and two women having a discussion **over** white wine.
- **TXT**: Ein Mann und zwei Frauen diskutieren **über** Weißwein.
- **IMG**: Ein Mann und zwei Frauen diskutieren **und trinnken** Weißwein.

Humans and MT systems

- **MT**: Ein Mann und zwei Frauen unterhalten sich **über** ein weißes Café.
- **MMT**: Ein Mann und zwei Frauen unterhalten sich **über** einen weißen Waschbecken.

Humans and MT systems

- **SRC**: A woman sitting on a very large **rock** smiling at the camera with trees in the background.
- **TXT**: Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen **Felsen** und lächelt in die Kamera.
- **IMG**: Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen **Stein** und lächelt in die Kamera.

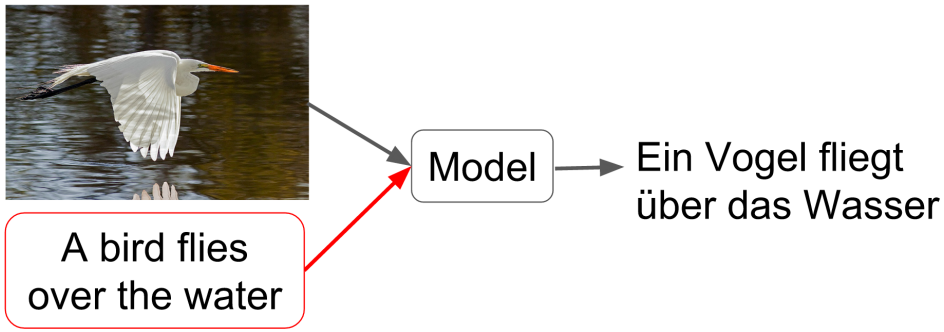
Humans and MT systems

- **MT**: Eine Frau sitzt auf einem sehr großen **Stein**, lächelt in die Kamera mit Bäumen im Hintergrund.
- **MMT**: Eine Frau sitzt auf einem sehr großen **Felsen** in die Kamera mit Bäumen im Hintergrund.

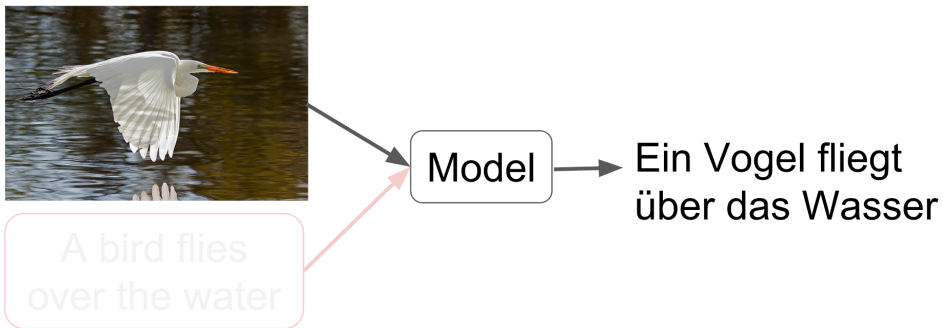
Overview

Related task

- **Crosslingual and Multilingual Image Description**: What can multilinguality bring to image description?
  - Take an image and generate a description in the target language, supported by the source language description



  - Take an image and generate a description in the target language; without source text (only training data in the source language)



Future directions

- More **realistic MT data** would not be descriptive, nor as simplistic, and would include cases where image is not related/relevant
- **Size of the dataset** needs to be significantly larger
- **Audio information** as additional modality for MMT: acoustic LDA, iVectors and acoustic embeddings [Deena et al., 2017]
- Current approaches surpassed common approaches using initialisation and double attention
- There is still a wide scope for exploration of the best **visual representation** and best **way to integrate** visual and textual information.

Multimodal Machine Translation

Lucia Specia

University of Sheffield

l.specia@sheffield.ac.uk

MTM - Lisbon, 1 Sept 2017

References I

Deena, S., Ng, R. W., Madhyashta, P., Specia, L., and Hain, T. (2017).

Semi-supervised adaptation of rnnlms by fine-tuning with domain-specific auxiliary features.

In Conference of the International Speech Communication Association, Stockholm, Sweden.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017).

Findings of the second shared task on multimodal machine translation and multilingual image description.

In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers, Copenhagen, Denmark. Association for Computational Linguistics.

Elliott, D., Frank, S., and Hasler, E. (2015).

Multi-language image description with neural sequence models.

CoRR, abs/1510.04709.

References II

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016).

Multi30k: Multilingual english-german image descriptions.

In 5th Workshop on Vision and Language, pages 70–74, Berlin, Germany.

Gella, S., Lapata, M., and Keller, F. (2016).

Unsupervised visual sense disambiguation for verbs using multimodal embeddings.

In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 182–192, San Diego, California.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015).

ImageNet Large Scale Visual Recognition Challenge.

International Journal of Computer Vision (IJCV), 115(3):211–252.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2014).

Imagenet large scale visual recognition challenge.

CoRR, abs/1409.0575.

References III

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015).

Show, attend and tell: Neural image caption generation with visual attention.

In ICML, volume 14, pages 77–81.