Grounded Sequence to Sequence Transduction

Lucia Specia

University of Sheffield, l.specia@sheffield.ac.uk

20 June, 2018

Schedule

Morning:

0900-0940	Multimodal learning (Lucia Specia)
0940-0945	Introduction to project (Lucia Specia)
0945-1045	Multimodal Machine Translation (Loïc Barrault)
1045-1100	Coffee Break
1100-1145	Multimodal ASR (Florian Metze)
1145-1215	Text Summarisation (Florian Metze)

Afternoon:

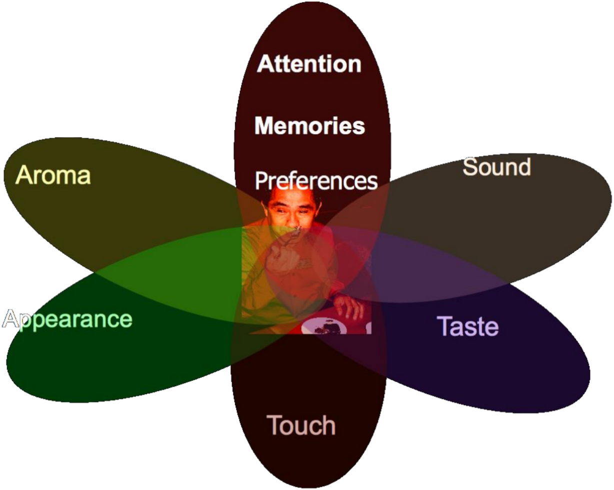
0130-0500	Lab on Multimodal MT and Summarisation

Sources

- Louis-Philippe Morency and Tadas Baltrusaitis. **Multimodal Machine Learning**, Tutorial at ACL 2017
- Desmond Elliott, Douwe Kiela and Angeliki Lazaridou. **Multimodal Learning and Reasoning**, Tutorial at ACL 2016

Multimodality

What do we mean?



Sensory modalities

What do we mean?

- **Modality**: type of information and/or the representation format in which information is stored.
- Examples:
  - **Natural language (spoken or written)**
  - Visual (from images or videos)
  - Auditory (voice, sounds and music)
  - Haptics / touch
  - Eye tracking
  - Other signals: electrocardiogram (ECG), infrared images, depth images, fMRI, etc.

What do we mean?

Verbal

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Vocal

- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual

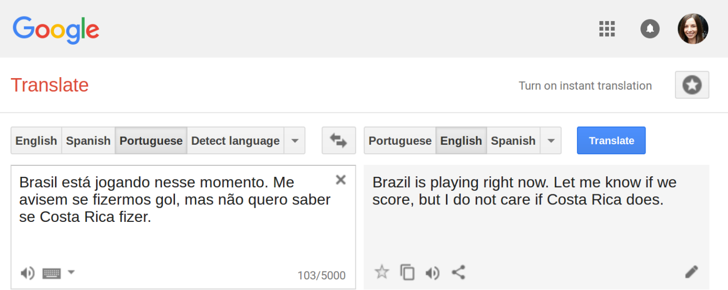
- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Why do we care?

Humans interact with the world in multimodal ways. Language understanding and generation is an not an exception.

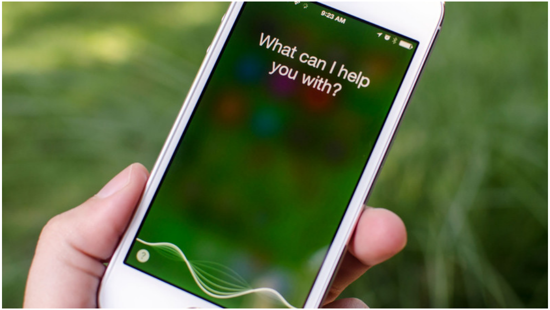
How well do we do?

NLP has advanced a lot:



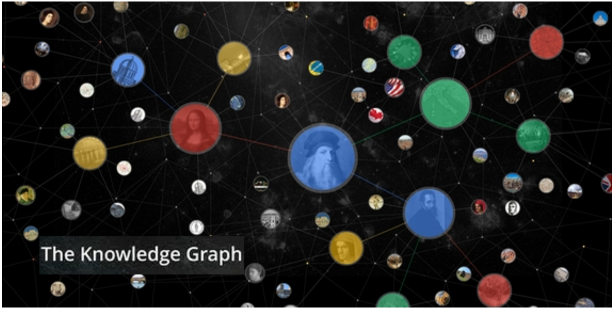
How well do we do?

NLP has advanced a lot:



How well do we do?

NLP has advanced a lot:

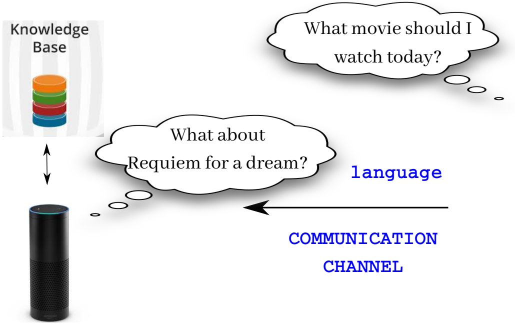


How well do we do?

However it is still mostly **monomodal** (speech or text):

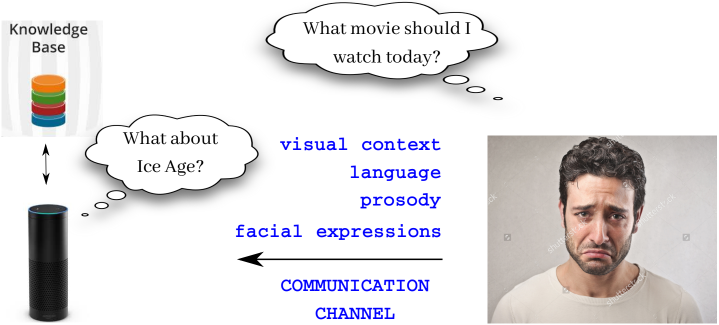
How well do we do?

However it is still mostly **monomodal** (speech or text):



How well do we do?

However it is still mostly **monomodal** (speech or text):

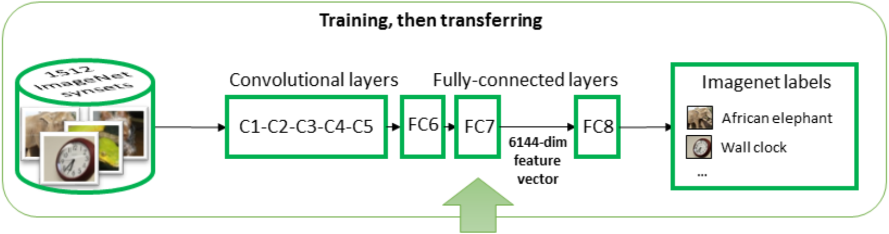


Moving beyond the linguistic modality

Some tasks are **inherently multimodal**. E.g. image captioning:

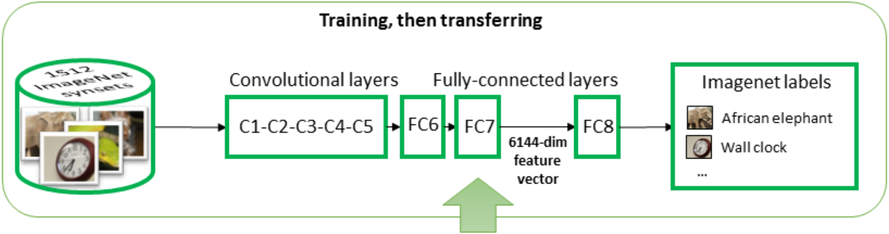
Girl in pink dress is jumping in air.

Image captioning



1. Train a **convolutional neural network** on a vision task e.g. AlexNet [Krizhevsky et al., 2012]
2. Do a **forward pass** given an image input
3. **Transfer** one or more layers (e.g. FC_7, or $CONV_5$) to an RNN to generate a description

Image captioning



1. Train a **convolutional neural network** on a vision task e.g. AlexNet [Krizhevsky et al., 2012]
2. Do a **forward pass** given an image input
3. **Transfer** one or more layers (e.g. FC_7, or $CONV_5$) to an RNN to generate a description

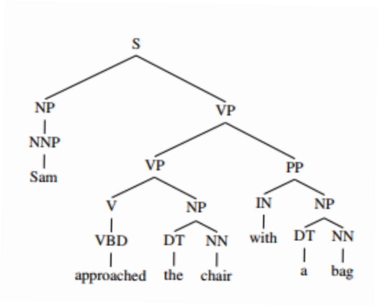
Others are not, e.g. Parsing, POS tagging, MT, Summarisation.

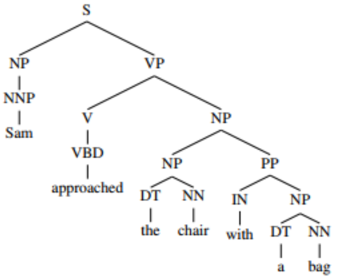
Multimodality helps with classic NLP tasks

PP attachment disambiguation [Berzak et al., 2015]

Sam approached the chair with a bag.

Sam approached the chair with a bag.





Multimodality helps with classic NLP tasks

PP attachment disambiguation [Berzak et al., 2015]

Sam approached the chair with a bag.

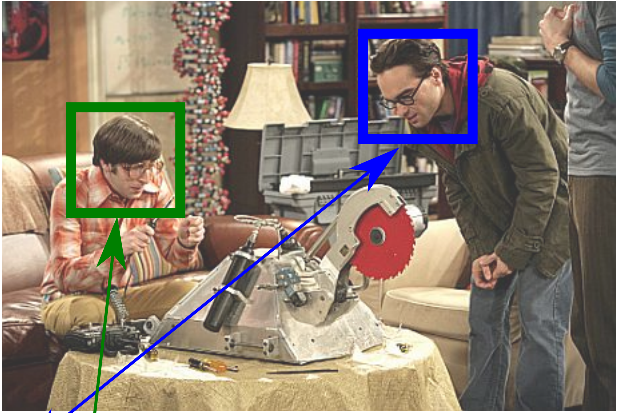
Multimodality helps with classic NLP tasks

Co-reference resolution [Ramanathan et al., 2014]

Leonard looks at the robot, while the only **engineer** in the room fixes it. **He** is amused.

Multimodality helps with classic NLP tasks

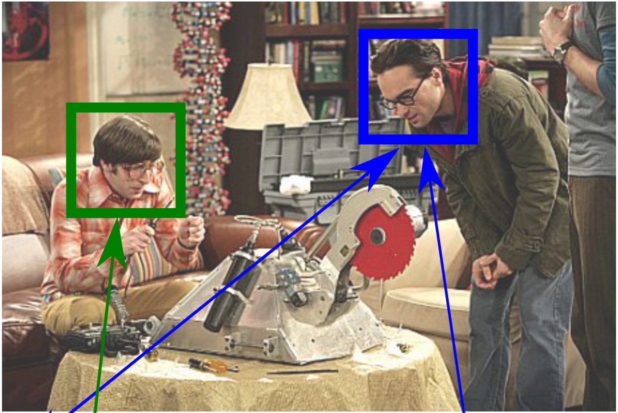
Co-reference resolution [Ramanathan et al., 2014]



Leonard looks at the robot, while the only **engineer** in the room fixes it. **He** is amused.

Multimodality helps with classic NLP tasks

Co-reference resolution [Ramanathan et al., 2014]



Leonard looks at the robot, while the only **engineer** in the room fixes it. **He** is amused.

Multimodality helps with classic NLP tasks

Machine translation [Frank et al., 2018]

- **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.

Multimodality helps with classic NLP tasks

Machine translation [Frank et al., 2018]

- **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.

Challenges

How do we do it?

Richer context models; better grounding

How do we do it?

Richer context models; better grounding

- Historical view by Morency & Baltrusaitis (2017):
  - The behavioral era (1970s until late 1980s)
  - The computational era (late 1980s until 2000)
  - The interaction era (2000–2010)
  - The deep learning era (2010s–) – focus on this lecture and our project

Core challenges in multimodal learning

Representation

Alignment

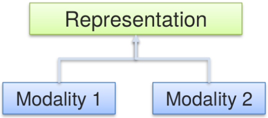
Fusion

Translation

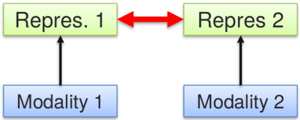
Co-learning

Learn how to represent and summarise multimodal data in a way that exploits the complementarity and redundancy.

- **Joint**:



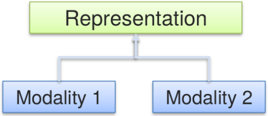
- **Coordinated**:



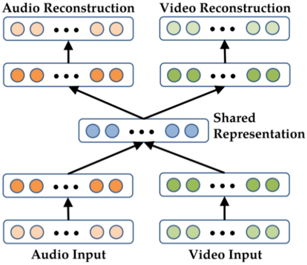
1: Representation

Learn how to represent and summarise multimodal data in a way that exploits the complementarity and redundancy.

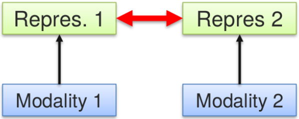
- **Joint**:



E.g. Multimodal autoencoders

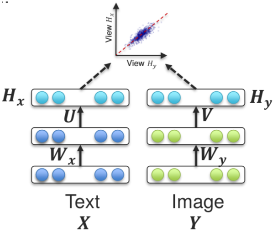


- **Coordinated**:



e.g. Deep CCA [Andrew et al., 2013]

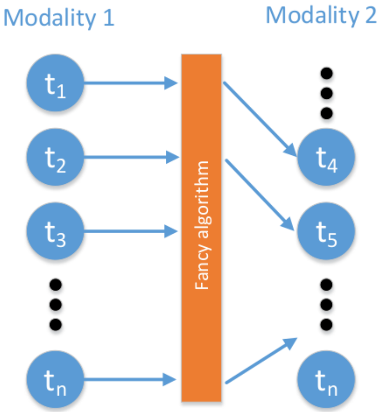
$$(u^*, v^*) = argmax_{corr}(u^T X, v^T Y)$$



2: Alignment

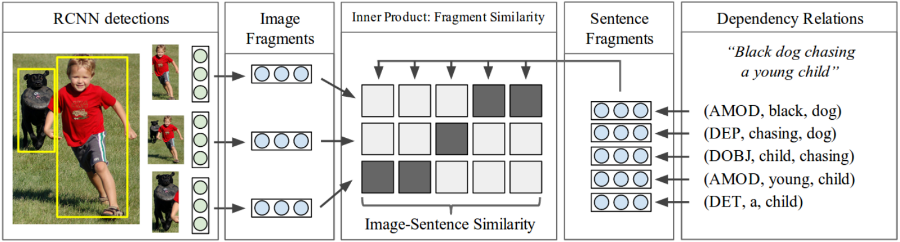
Identify the direct relations between elements from different modalities.

- **Explicit**: Directly find correspondences between elements of different modalities.
- **Implicit**: Indirectly uses latent alignment of modalities.



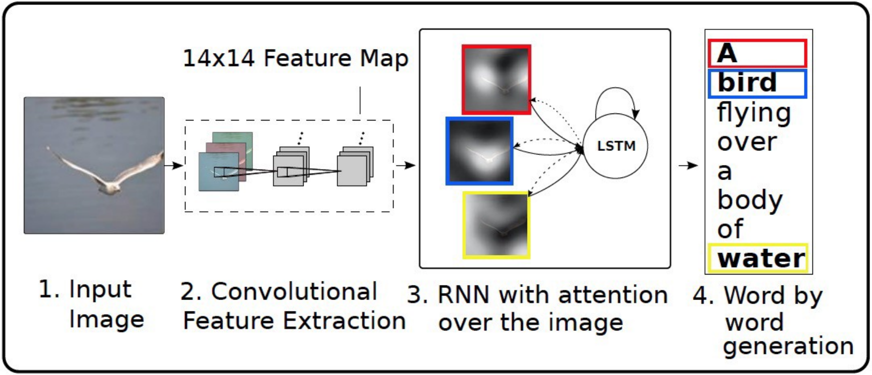
2: Alignment

Explicit: Learn to associate fragments in images and sentence descriptions [Karpathy et al., 2014].



2: Alignment

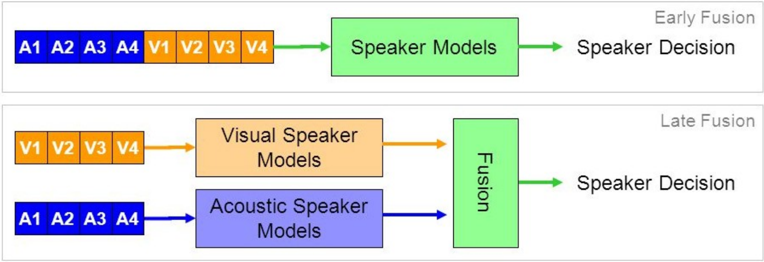
Implicit: Attention mechanism in image captioning [Xu et al., 2015].



3: Fusion

How/when to join information from various modalities.

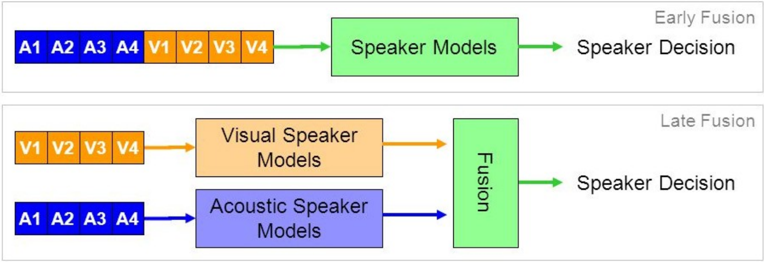
- **Early fusion**:
- **Late fusion**:



3: Fusion

How/when to join information from various modalities.

- **Early fusion**:
- **Late fusion**:



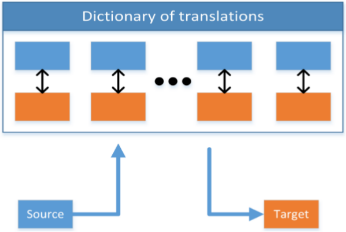
Fusion can model agnostic: e.g. feature fusion or ensemble via voting.

Image from: `http://images.slideplayer.com/11/3284040/slides/slide_3.jpg`

4: Translation

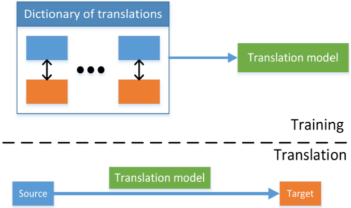
Given an entity in one modality, how to generate the same entity in a different modality.

- **Example-based**:

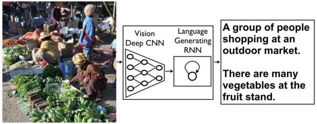


E.g. KNN in query-based image retrieval

- **Model-based**:



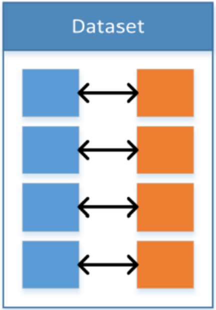
E.g. [Vinyals et al., 2015]



5: Co-learning

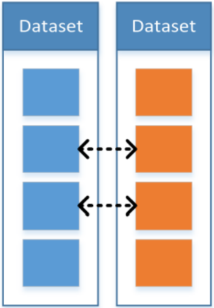
How to transfer knowledge between modalities, including representations and models.

- **Parallel**:



E.g. co-training, multi-task learning

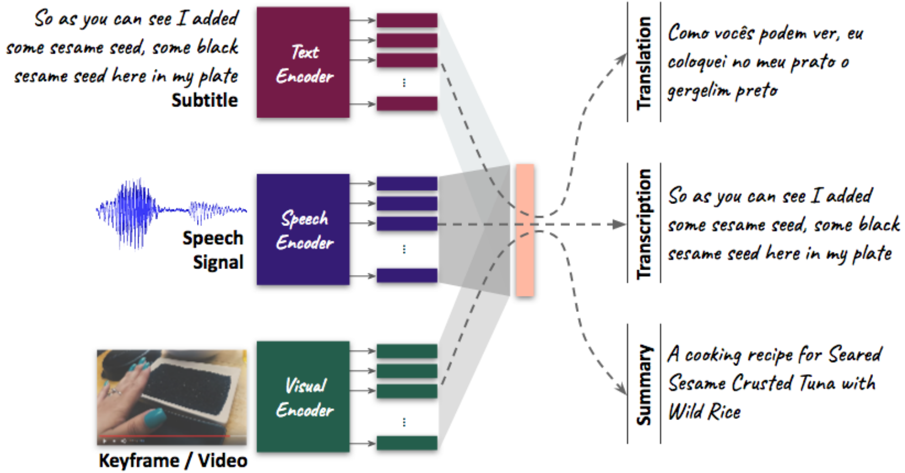
- **Non-parallel**:



E.g. Zero-shot learning

Our project

Grounded Seq2Seq Transduction



https://srvk.github.io/jsalt-2018-grounded-s2s/

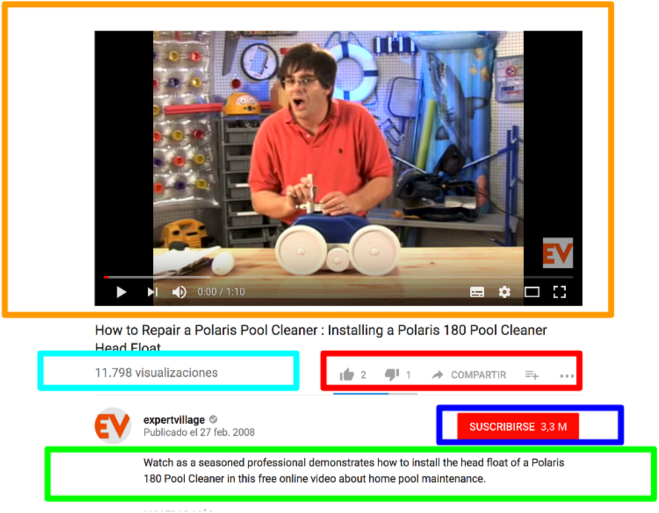
Grounded Seq2Seq Transduction

- Core method: **sequence-to-sequence models**
- Challenges to address:
  - From bimodal to multimodal
  - Different tasks: MT, ASR, Summarisation (plus auxiliary)
  - Joint learning (multi-task learning)
  - Move closer to video understanding
  - Explore temporal nature of videos
  - Understand which modalities help for which tasks
  - Datasets

Grounded Seq2Seq Transduction

- **Dataset**: largest multimodal, bilingual (will be multilingual) dataset
    - 2,000 hours of how-to videos
    - Video, audio, human transcripts, 'summary'
    - Wide range of topics
    - 480 hours of how-to videos with translations
    - 19,000 short videos
    - 160,000 segment translations (100,000 thus far)

Dataset


0:00 / 1:10
How to Repair a Polaris Pool Cleaner : Installing a Polaris 180 Pool Cleaner Head Float
11.798 visualizaciones
2
1
COMPARTIR
expertvillage
Publicado el 27 feb. 2008
SUSCRIBIRSE 3,3 M
Watch as a seasoned professional demonstrates how to install the head float of a Polaris 180 Pool Cleaner in this free online video about home pool maintenance.

Schedule

Morning:

0900-0940 Multimodal learning (Lucia Specia)

0940-0945 Introduction to project (Lucia Specia)

0945-1045 Multimodal Machine Translation (Loïc Barrault)

1045-1100 Coffee Break

1100-1145 Multimodal ASR (Florian Metze)

1145-1215 Text Summarisation (Florian Metze)

Afternoon:

0130-0500 Lab on Multimodal MT and Summarisation

References I

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013).
Deep canonical correlation analysis.
In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1247–III–1255. JMLR.org.

Berzak, Y., Barbu, A., Harari, D., Katz, B., and Ullman, S. (2015).
Do you see what i mean? visual resolution of linguistic ambiguities.
Conference on Empirical Methods in Natural Language Processing (EMNLP).

Frank, S., Elliott, D., and Specia, L. (2018).
Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices.
Natural Language Engineering, 24(3):393–413.

Karpathy, A., Joulin, A., and Fei-Fei, L. (2014).
Deep fragment embeddings for bidirectional image sentence mapping.
In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 1889–1897, Cambridge, MA, USA. MIT Press.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Ramanathan, V., Joulin, A., Liang, P., and Fei-Fei, L. (2014).
Linking people with "their" names using coreference resolution.
In *European Conference on Computer Vision (ECCV)*.

References II

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015).
Show and tell: A neural image caption generator.
pages 3156–3164.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015).
Show, attend and tell: Neural image caption generation with visual attention.
In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.