MULTILINGUAL IMAGE DESCRIPTION WITH NEURAL SEQUENCE MODELS

Desmond Elliott, Stella Frank, Eva Hasler

February 3, 2016

iV&L Net Working Group 3 Meeting

Institute for Logic, Language and Computation

Stella Frank

Eva Hasler

Grounded Semantics [Silberer and Lapata, 2014]

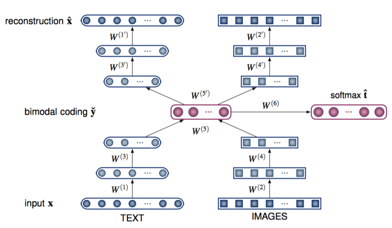
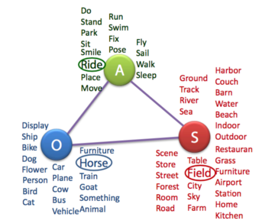
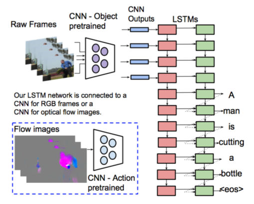


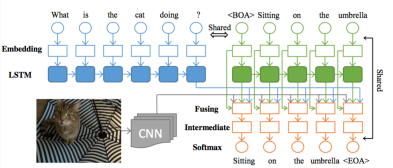
Image Description [Farhadi et al., 2010]



Video Description [Venugopalan et al., 2015]



Question-Answering [Gao et al., 2015]



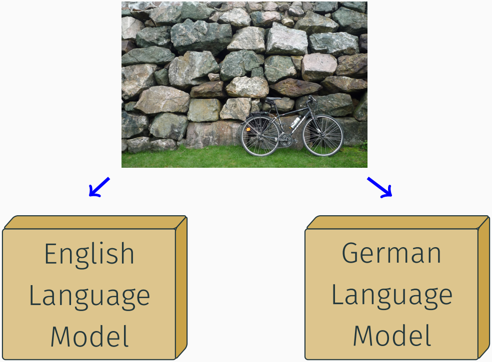
Language Generation Model

A bike is leaning against a stone wall

See Bernardi et al. [2016] for an overview of datasets, models, and evaluations.

THIS TALK: DESCRIBING IMAGES IN MULTIPLE LANGUAGES

- Extend image description generation to new languages
- Text-based image search in any language
- Localised alt-text generation on the Web
- Translate movie descriptions



Ein Rad steht neben der Mauer $\longrightarrow$ A bicycle / wheel $\cdots$

English Language Model

German Language Model

HOW CAN WE EXPLOIT MULTILINGUAL MULTIMODAL CONTEXT?

Ein Rad steht neben der Mauer $\longrightarrow$ A bicycle / wheel $\cdots$

<Image features> $\longrightarrow$ A ? is leaning against the wall

English Language Model

German Language Model

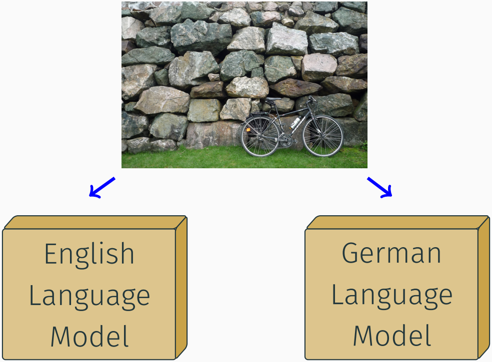
HOW CAN WE EXPLOIT MULTILINGUAL MULTIMODAL CONTEXT?

Ein Rad steht neben der Mauer $\longrightarrow$ A bicycle / wheel $\cdots$

<Image features> $\longrightarrow$ A ? is leaning against the wall

Possible solutions:

- Collect more data



Ein Rad steht neben der Mauer $\longrightarrow$ A bicycle / wheel $\cdots$

<Image features> $\longrightarrow$ A ? is leaning against the wall

Possible solutions:

- Collect more data
- Exploit data in a different modality (images or video)

$\swarrow$ $\searrow$

English Language Model

German Language Model

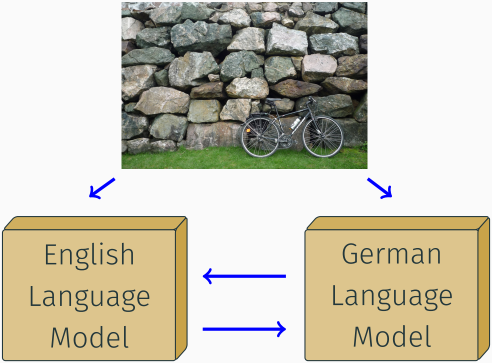
HOW CAN WE EXPLOIT MULTILINGUAL MULTIMODAL CONTEXT?

Ein Rad steht neben der Mauer ⟶ A bicycle / wheel ···

<Image features> ⟶ A ? is leaning against the wall

Possible solutions:

- Collect more data
- Exploit data in a different modality (images or video)



Let t be the **target** language description, s be the **source** language description and i be the **image**.

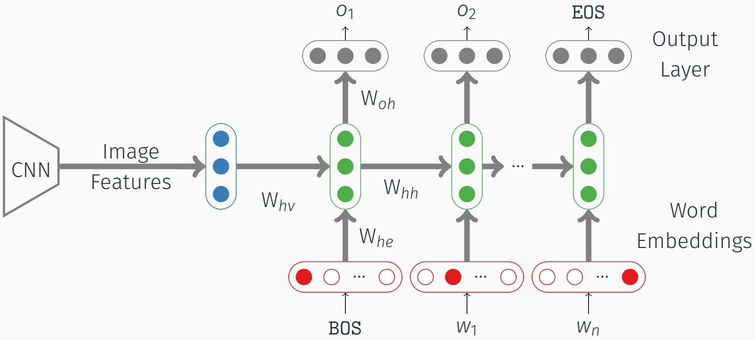
1. Multimodal Machine Translation
    - Always given a source description and image
    - $\hat{t} = \text{argmax}_t \, p(t|i, s)$

Let t be the **target** language description, s be the **source** language description and i be the **image**.

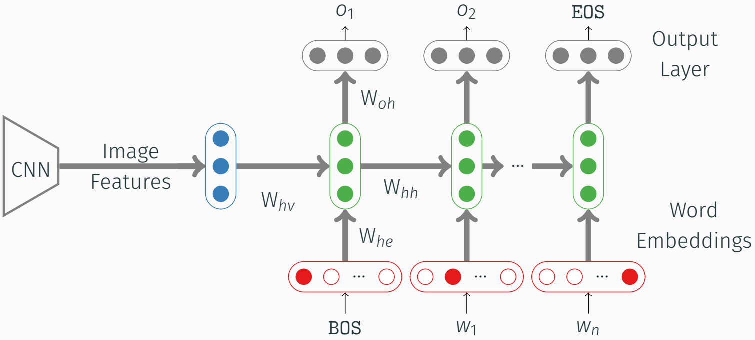
1. Multimodal Machine Translation
   - Always given a source description and image
   - $\hat{t} = \text{argmax}_t \, p(t|i, s)$
2. Crosslingual Image Description
   - Automatically generate a source language description
   - $\hat{t} = \text{argmax}_t \, p(t|i, \hat{s})$

MULTILINGUAL MULTIMODAL MODEL

MULTIMODAL LANGUAGE MODELS [VINYALS ET AL., 2015, KARPATHY AND FEI-FEI, 2015]

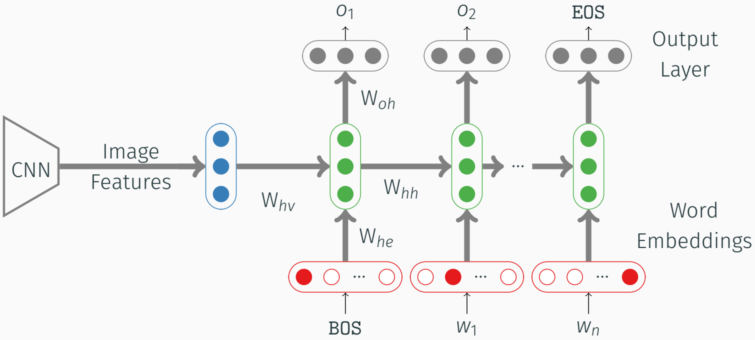


[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



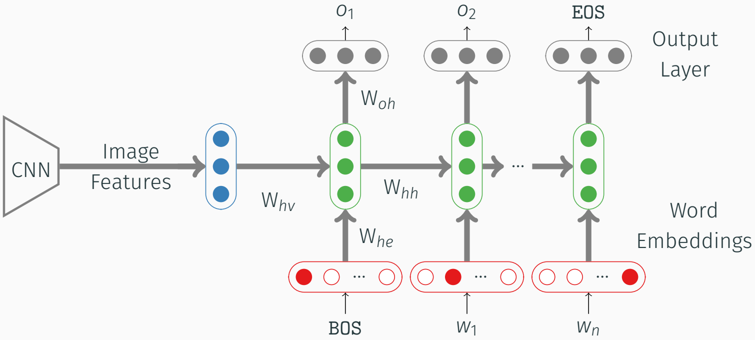
- $e_i = W_{he} w_i$

Multimodal Language Models [Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



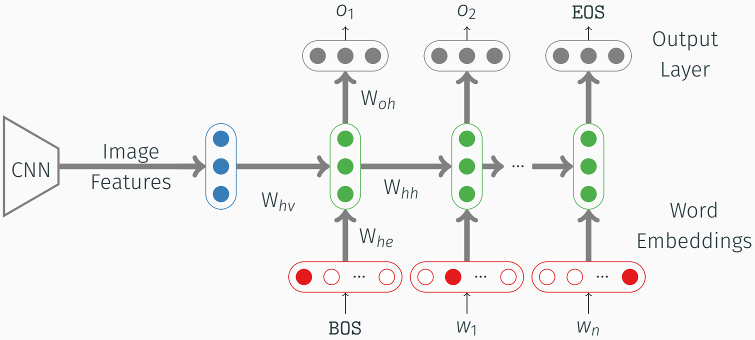
- $e_i = W_{he} w_i$
- $h_i = f(W_{hh} h_{i-1} + e_i + \mathbb{1}(t = 0) \cdot W_{hv} v)$

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



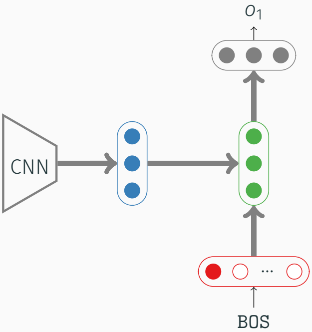
- $e_i = W_{he} w_i$
- $h_i = f(W_{hh} h_{i-1} + e_i + \mathbb{1}(t = 0) \cdot W_{hv} v)$
- $o_i = \text{softmax}(W_{oh} h_i)$

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



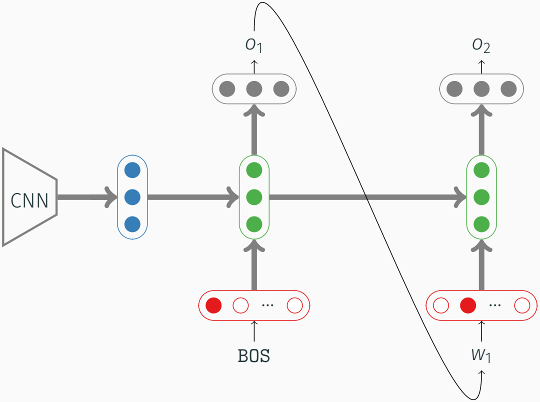
- $e_i = W_{he} w_i$
- $h_i = f(W_{hh} h_{i-1} + e_i + \mathbb{1}(t=0) \cdot W_{hv} v)$
- $o_i = \text{softmax}(W_{oh} h_i)$
- $Loss = -\sum_{n=1}^{N} \sum_{i=1}^{K} log\ p(o_i)$

INFERENCE WITH MULTIMODAL LANGUAGE MODELS [Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



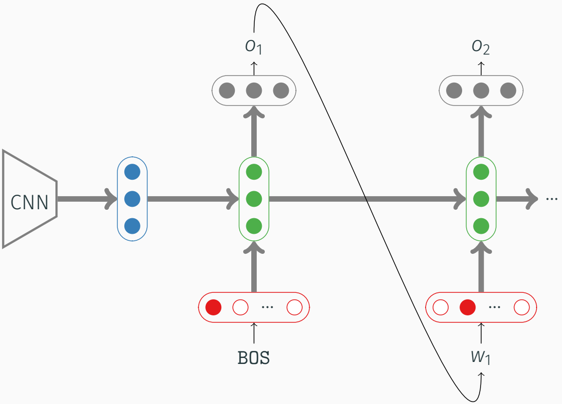
- Initialise with image features and BOS token

INFERENCE WITH MULTIMODAL LANGUAGE MODELS [Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



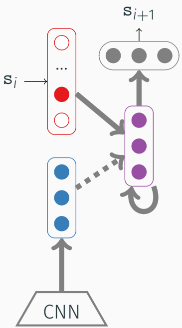
- Initialise with image features and BOS token
- Feed sampled word into the next timestep

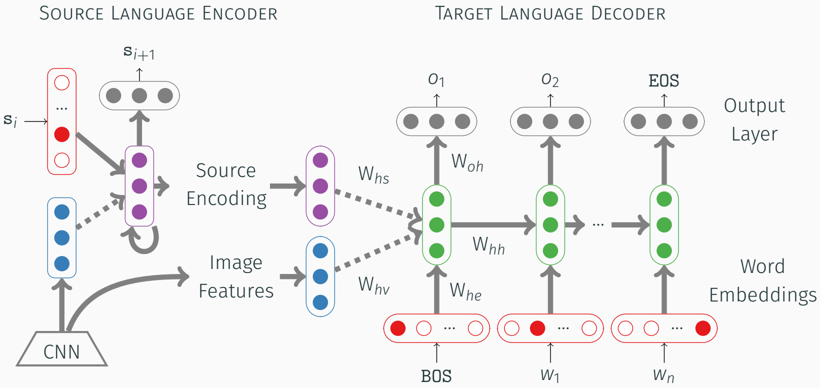
INFERENCE WITH MULTIMODAL LANGUAGE MODELS [Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



- Initialise with image features and BOS token
- Feed sampled word into the next timestep
- Decode until emit EOS token

SOURCE LANGUAGE ENCODER





$$h_i = f(W_{hh}h_{-1} + e_i + \mathbb{1}(t = 0) \cdot W_{hv}v + \mathbb{1}(t = 0) \cdot W_{hs}s)$$

- Each model trained towards its own objective, unlike Sequence-to-Sequence Learning [Sutskever et al., 2014]
    - CNN: object recognition
    - Source LM: source language generation
    - Target LM: target language generation
- MMLM learns task-specific representations given transferred inputs
    - e.g. Target-LM with **multimodal** source features vs. separate **visual and source** features
    - Easily work on new languages with fixed input representations

EXPERIMENTS

- Generate description in target language
- Measures[1]: Meteor, BLEU, Perplexity

1. Multimodal Machine Translation
    - Always given a source description and image
2. Crosslingual Image Description
    - Given an image, automatically generate a source description with a source MLM
        - and pass encoded textual features to a target LM
        - and pass encoded visual+textual features to a target LM

[1] See Elliott and Keller [2014] and Vedantam et al. [2015] for more details on measuring image description quality

IAPR-TC12 DATASET [GRUBINGER ET AL., 2006]

1. a yellow building with white columns in the background

2. two palm trees in front of the house

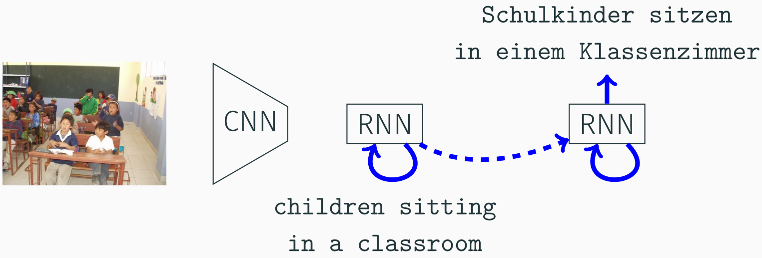
1. ein gelbes Gebäude mit weißen Säulen im Hintergrund

2. zwei Palmen vor dem Haus

- 17,655 training / 1,962 testing
- Up to five semantically diverse descriptions / image
  - We use **only** the first description
- Descriptions translated from English to German

- Models are built using Keras library
- Adam optimiser [Kingma and Ba, 2014]
- Mini-batches of 100 examples
- Dropout over word, visual, and source features ($p = 0.5$)
- LSTM with 256-D memory cell [Hochreiter and Schmidhuber, 1997]
- 4096-D visual features from 15th layer of VGG-16 CNN [Simonyan and Zisserman, 2015]
- 256-D source feature vectors
- 256-D word embedding features
- Vocabulary size German: 2,374, English: 1,763 (UNK$<$3)

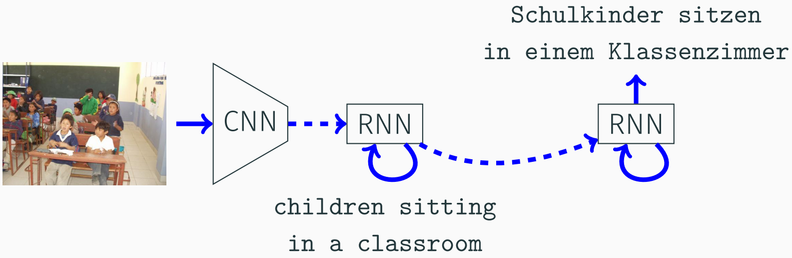
TARGET MODEL



SOURCE MODEL

TARGET MODEL

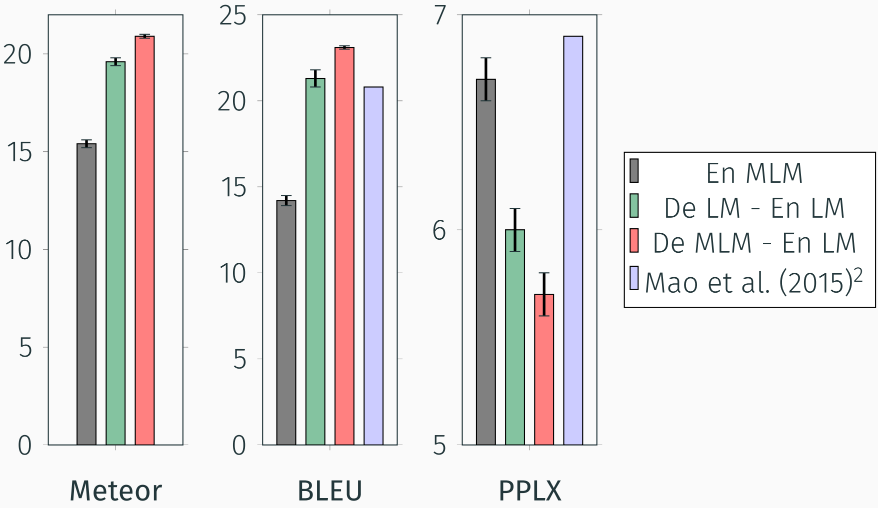
MODELS: SOURCE MLM → TARGET LM



SOURCE MODEL

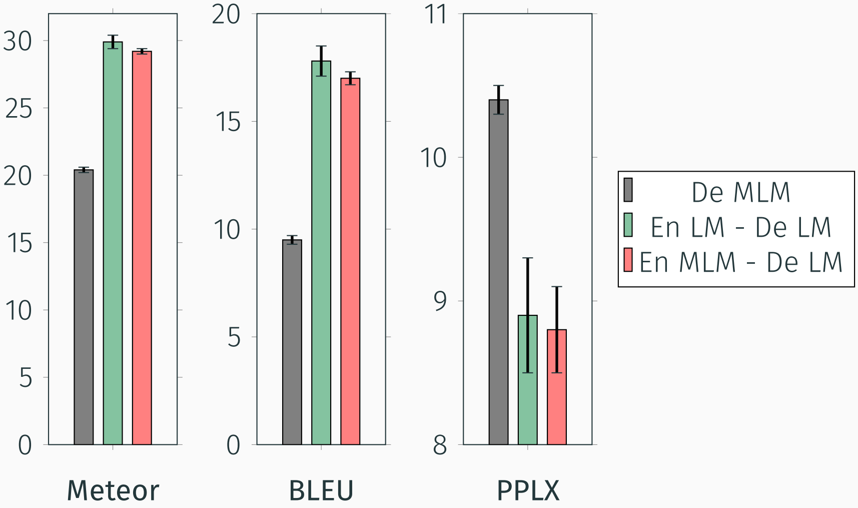
TARGET MODEL

RESULTS

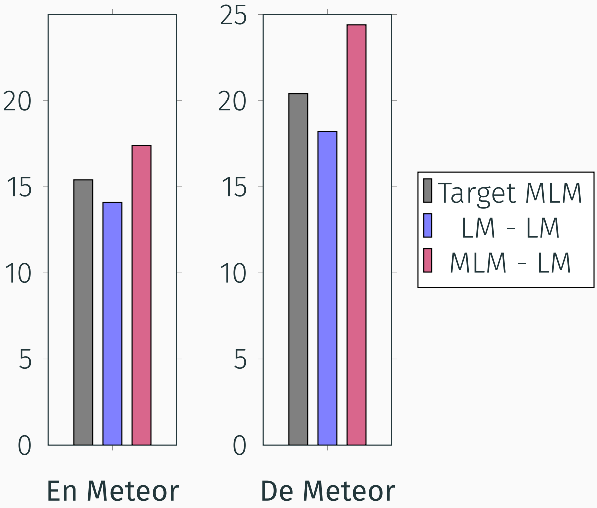


[2] Trained and evaluated over all references.

MULTIMODAL TRANSLATION: GERMAN RESULTS[3]



[3] First non-English image description results.



Source descriptions automatically generated by Source-MLM

MLM: a man is standing on a grey rock in the foreground

De Ref: bergsteiger klettern auf einen sehr steilen eishang

MLM-MLM: tourists are climbing up a snowy slope

[4] Thousands of examples from all models at
http://staff.fnwi.uva.nl/d.elliott/GroundedTranslation/

De MLM:
ein mann und eine frau stehen an einem sandstrand mit dem meer im hintergrund

En MLM: a man with a black jacket and a black jacket is standing in a brown rocky desert landscape

En MLM-LM:
a man and a woman are standing in a reed boat on a lake

- t-SNE plots of the LSTM memory cell at t=0
- MLM $\rightarrow$ MLM: closer to pictures of snow!

(a) En MLM

(b) De MLM $\rightarrow$ En MLM

- How well does this generalise to other languages?
- Attention-based Image Description [Xu et al., 2015]
- Compare with target-side translation retrieval with multimodal features [Hitschler and Riezler, 2016]
- Human judgements of generated descriptions
- Larger datasets (Shared Task at WMT16!)
- Multilingual video description, other tasks . . .

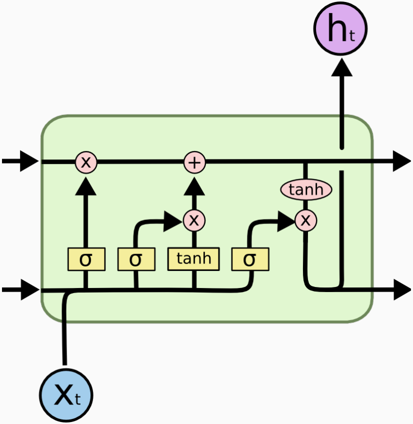
- Multilingual Image Description is a natural extension of Image Description
- MMLM transfers multimodal features between languages
- Transferring multilingual multimodal representations between languages improves image description quality
- Code: `http://github.com/elliottd/GroundedTranslation`

APPENDICES

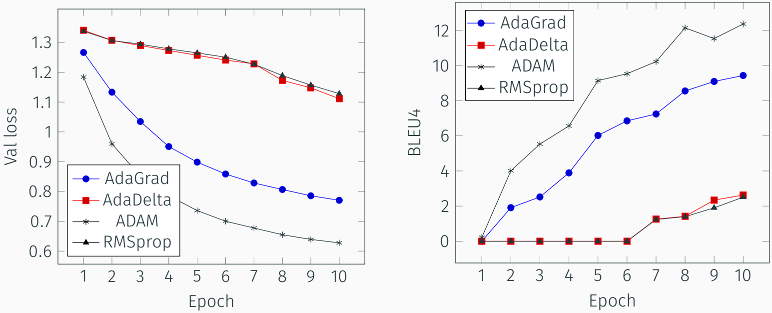
	BLEU4	Meteor	PPLX
En MLM	14.2 ± 0.3	15.4 ± 0.2	6.7 ± 0.0
De LM $\rightarrow$ En LM	21.3 ± 0.5	19.6 ± 0.2	6.0 ± 0.1
Mao et al. [2015]	20.8	—	6.92
De MLM $\rightarrow$ En MLM	18.0 ± 0.3	18.0 ± 0.2	6.3 ± 0.1
De LM $\rightarrow$ En MLM	17.3 ± 0.5	17.6 ± 0.5	6.3 ± 0.0
De MLM $\rightarrow$ En LM	$\mathbf{23.1} \pm \mathbf{0.1}$	$\mathbf{20.9} \pm \mathbf{0.0}$	5.7 ± 0.1

	BLEU4	Meteor	PPLX
De MLM	9.5 ± 0.2	20.4 ± 0.2	10.35 ± 0.1
En LM $\rightarrow$ De LM	**17.8 ± 0.7**	**29.9 ± 0.5**	8.95 ± 0.4
En MLM $\rightarrow$ De MLM	11.4 ± 0.7	23.2 ± 0.9	9.69 ± 0.1
En LM $\rightarrow$ De MLM	12.1 ± 0.5	24.0 ± 0.3	10.2 ± 0.7
En MLM $\rightarrow$ De LM	**17.0 ± 0.3**	**29.2 ± 0.2**	8.84 ± 0.3

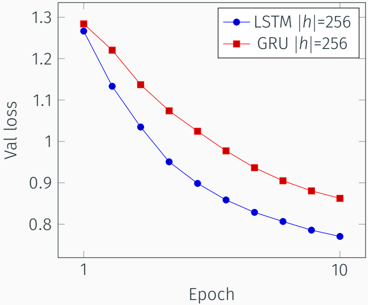
RNN ARCHITECTURE: LONG-SHORT TERM MEMORY [HOCHREITER AND SCHMIDHUBER, 1997]

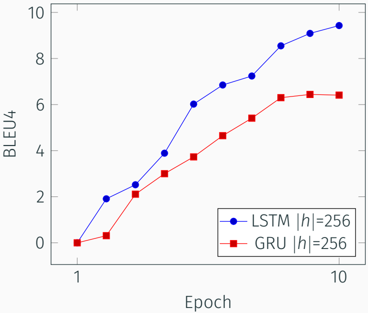


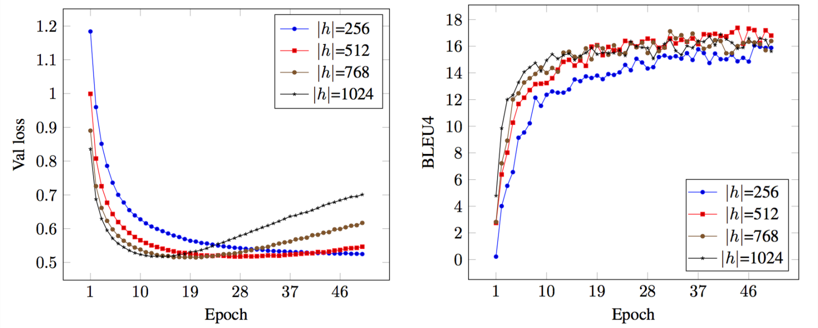
Credit: Christopher Olah


Val loss
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
1 2 3 4 5 6 7 8 9 10
Epoch
AdaGrad
AdaDelta
ADAM
RMSprop
BLEU4
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10
Epoch
AdaGrad
AdaDelta
ADAM
RMSprop

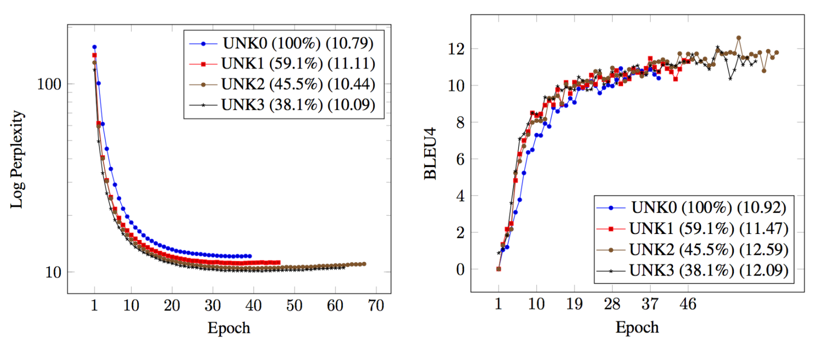
EFFECT OF RNN TYPE

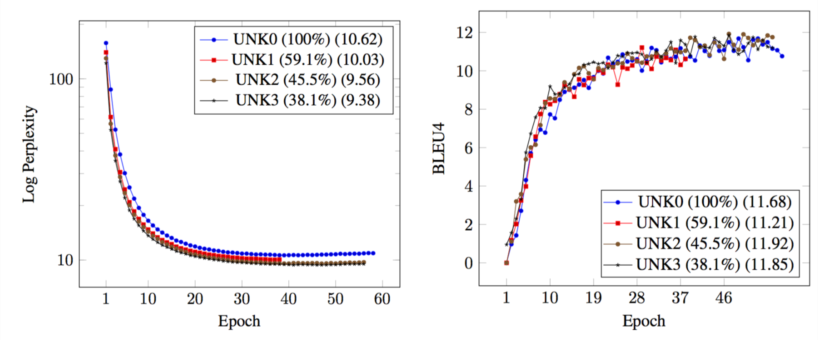


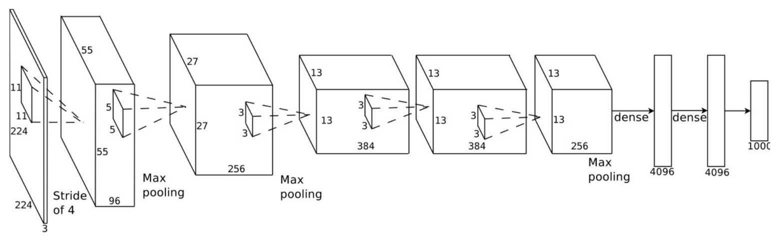



Val loss
Epoch
|h|=256
|h|=512
|h|=768
|h|=1024
BLEU4
Epoch
|h|=256
|h|=512
|h|=768
|h|=1024

EFFECT OF DECOMPOUNDING GERMAN WORDS



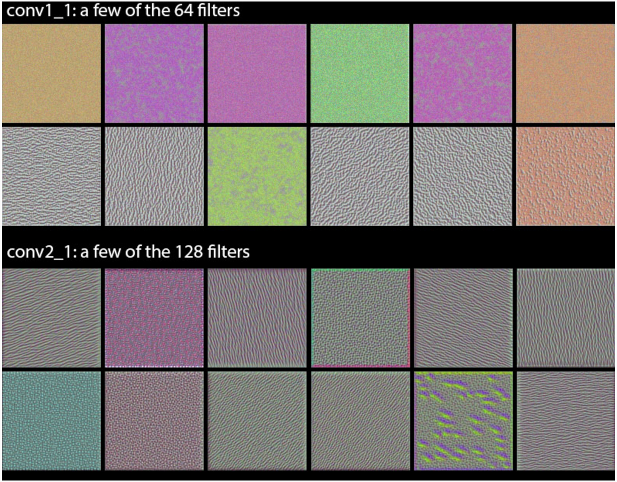

UNK0 (100%) (10.62)
UNK1 (59.1%) (10.03)
UNK2 (45.5%) (9.56)
UNK3 (38.1%) (9.38)
Log Perplexity
100
10
1
10
20
30
40
50
60
Epoch
12
10
8
6
4
2
0
BLEU4
UNK0 (100%) (11.68)
UNK1 (59.1%) (11.21)
UNK2 (45.5%) (11.92)
UNK3 (38.1%) (11.85)
1
10
19
28
37
46
Epoch



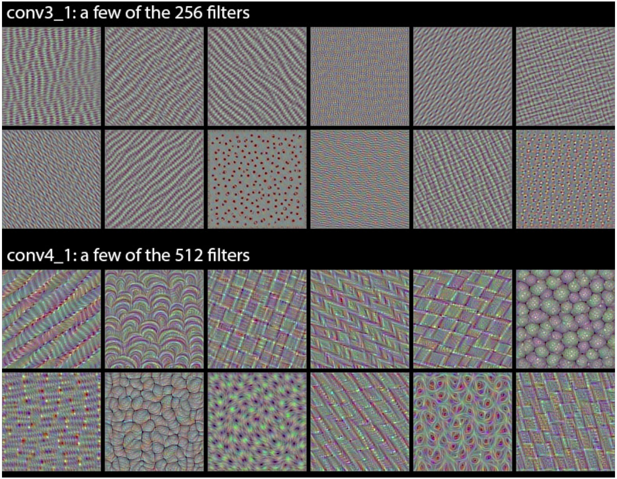
Credit: Alex Krizhevsky

- Trained to predict 1000 object labels
- Over 1m training images
- Visual features transferred from the penultimate layer

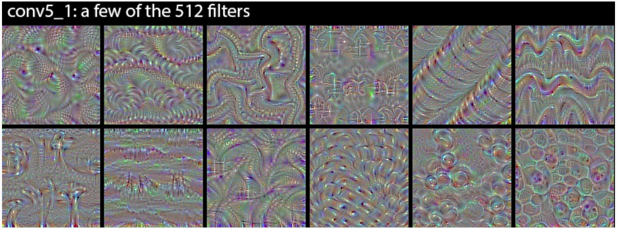
VISUALISING CNN FILTERS



Credit: François Chollet



Credit: François Chollet



Credit: François Chollet

Raffaella Bernardi, Ruken Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *To appear in JAIR*, 2016.

D Elliott, S Frank, and E Hasler. Multilingual image description with neural sequence models. *CoRR*, abs/1510.04709, 2015.

Desmond Elliott and Frank Keller. Comparing Automatic Evaluation Measures for Image Description. In *ACL*, 2014.

Ali Farhadi, M Hejrati, Mohammad Amin Sadeghi, P Young, C Rashtchian, J Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015.

M. Grubinger, P. D. Clough, H. Muller, and D. Thomas. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *LREC*, 2006.

J. Hitschler and S. Riezler. Multimodal Pivots for Image Caption Translation. *CoRR*, January 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv*, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). *ICLR*, 2015.

Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR '15*, 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence - video to text. In *ICCV*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.