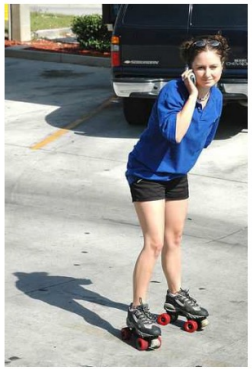
COMP 646: Deep Learning for Vision and Language

Multimodal Machine Translation

Today's Class

- Brief overview on
  - Multilingual Image Captioning
  - Multimodal Machine Translation

Multilingual Image Captioning



1. There is a young girl on her cell-phone while skating.
2. Eine Frau im blauen Shirt telefoniert beim Rollschuhfahren.

(b) Independent descriptions

In the latest version

- Captions in English and German
- 30,000 images
- 5 captions per image per language = 5 * 30,000 * 2
- Images from the Flickr30k dataset

Multi30K: Multilingual English-German Image Descriptions

Desmond Elliott and **Stella Frank** and **Khalil Sima'an**
ILLC, University of Amsterdam
{d.elliott, s.c.frank, k.simaan}@uva.nl

Lucia Specia
Sheffield University
l.specia@sheffield.ac.uk

https://arxiv.org/abs/1605.00459

Image Captioning Models

Vision
Deep CNN

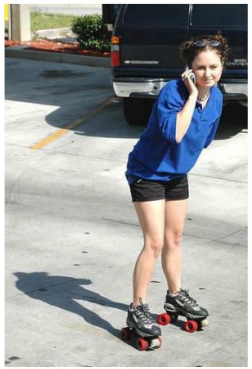
Recurrent
Text Decoder

Caption in German

Recurrent
Text Decoder

Caption in English

Multimodal Machine Translation



1. Trendy girl talking on her cellphone while gliding slowly down the street
2. Ein schickes Mädchen spricht mit dem Handy während sie langsam die Straße entlangschwebt.

(a) Translations

In the latest version

- Captions in English, German, French and Czech
- 30,000 images
- 1 captions per image per language = 30,000 * 4
- Images from the Flickr30k dataset

Multi30K: Multilingual English-German Image Descriptions

Desmond Elliott and **Stella Frank** and **Khalil Sima'an**
ILLC, University of Amsterdam
{d.elliott, s.c.frank, k.simaan}@uva.nl

Lucia Specia
Sheffield University
l.specia@sheffield.ac.uk

https://arxiv.org/abs/1605.00459

Multimodal Machine Translation

Captions

- 紅白で統一されたスタイリッシュなキッチン
- 白黒の床に置かれた赤と白に統一されたキッチン
- 赤と白と黒で統一されたキッチン
- モノトーンと赤で統一されたモダンなキッチン
- キッチンには銀色の取っ手がついた赤色の収納庫がある

STAIR Captions:
Constructing a Large-Scale Japanese Image Caption Dataset

Yuya Yoshikawa **Yutaro Shigeto** **Akikazu Takeuchi**
Software Technology and Artificial Intelligence Research Laboratory (STAIR Lab)
Chiba Institute of Technology
2-17-1, Tsudanuma, Narashino, Chiba, Japan.
{yoshikawa,shigeto,takeuchi}@stair.center

- Based on the COCO Dataset but with Japanese – so combined with COCO it is 10 captions per image for > 100,000 images.

https://arxiv.org/abs/1705.00823

Integrating Vision and Language: Multimodal Machine Translation

P(caption' / image, caption)

\+ **Two people playing with a kite on the beach** ⟶ **Dos personas jugando en la playa con una cometa.**

Datasets: Multi30K: English-German, English-French, IAPR-TC 12 English-German, Pascal Sentence Japanese-English

Grubinger et al 2006, Specia et al 2016, Elliott et al 2017, Calixto et al 2016, 2017a, 2017b, Caglayan et al 2017, Helcl et al 2017.

Integrating Vision and Language: Multimodal Machine Translation

P(caption' / image, caption)

Vision
Deep CNN

Recurrent
Text Decoder

Dos personas jugando en la playa con una cometa.

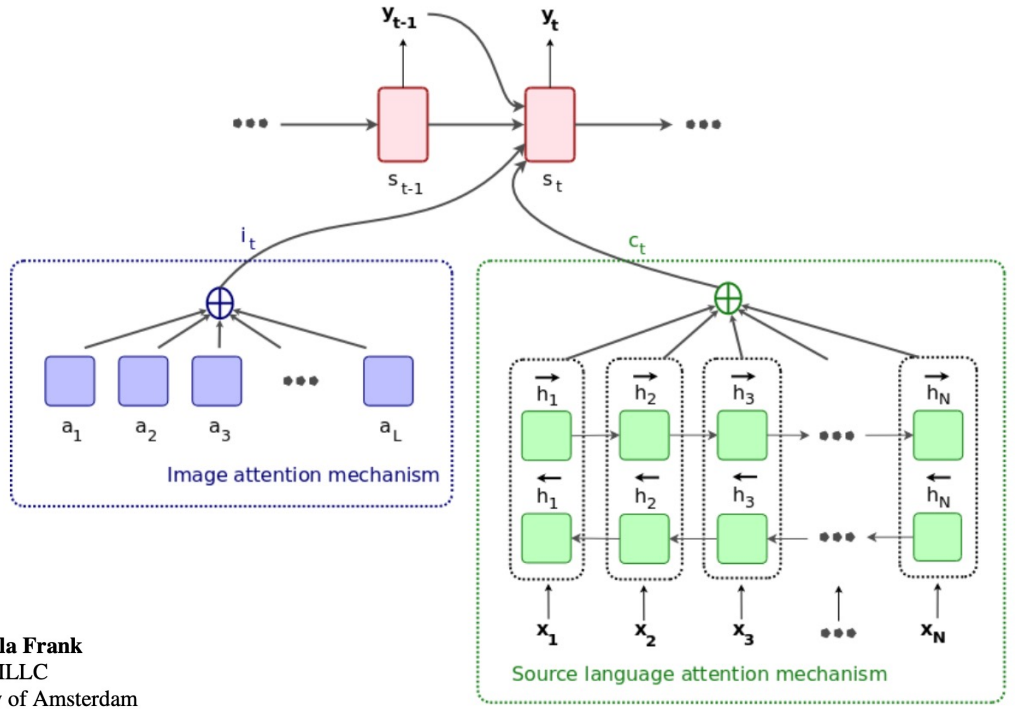
Recurrent
Text Encoder

Two people playing with a kite on the beach

e.g. GRU Encoder + CNN Encoder + GRU Decoder (Attention)

Caglayan et al 2017, Calixto et al 2016

Sample Multimodal MT Model in more detail



DCU-UvA Multimodal MT System Report

Iacer Calixto
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
iacer.calixto@adaptcentre.ie

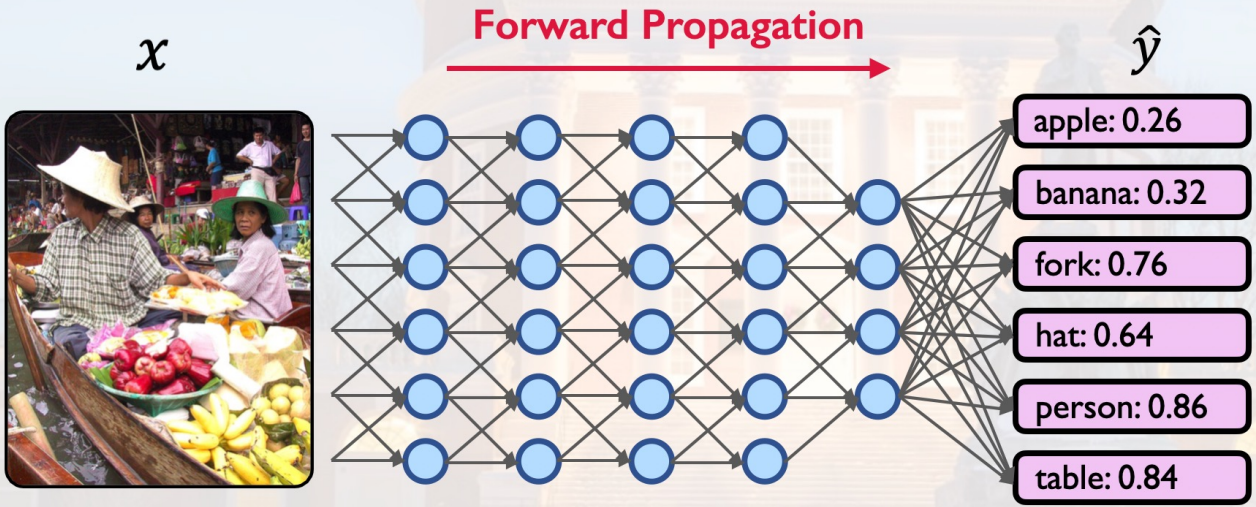
Desmond Elliott
ILLC
University of Amsterdam
Science Park 107
1098 XG Amsterdam
d.elliott@uva.nl

Stella Frank
ILLC
University of Amsterdam
Science Park 107
1098 XG Amsterdam
s.c.frank@uva.nl

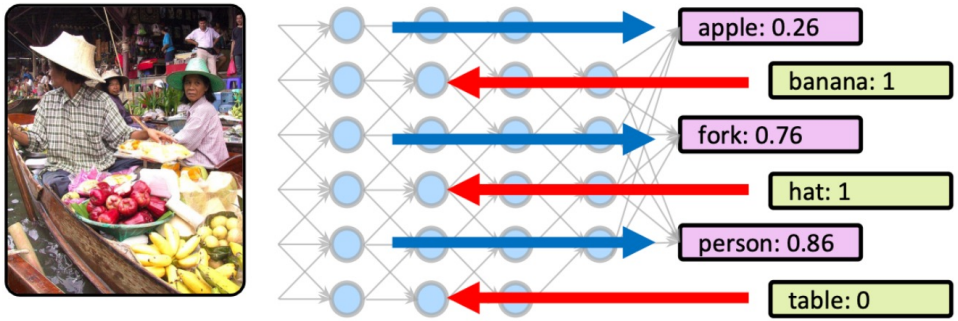
https://www.aclweb.org/anthology/W16-2359.pdf

Deep Neural Networks are quite Rigid

[In most cases] once a model its trained, **input** and **output** variables are **fixed**.



Feedback Propagation Inference as a solution

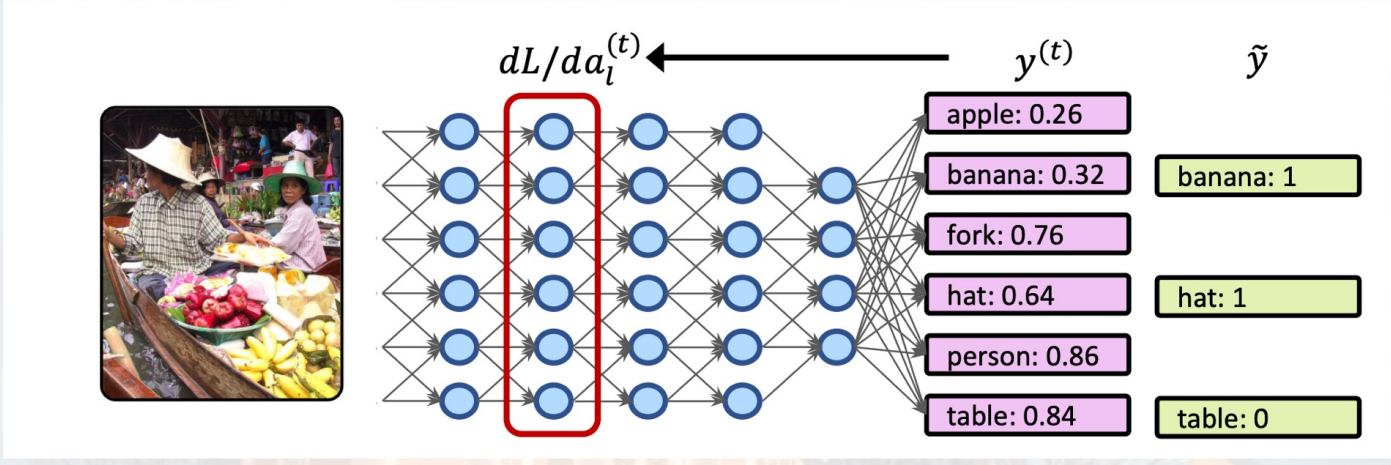


We found that output variables can be used as input at test time by iterative inference using backward and forward passes on intermediate features.

$$a_l^* = \text{argmin}_{a_l} L(Y_\text{k}, F_\text{k}^{(l)}(a_l, \Theta)),$$
$$\hat{Y}_\text{u} = F_\text{u}^{(l)}(a_l^*, \Theta).$$

Feedback-prop: Convolutional Neural Network Inference under Partial Evidence
Tianlu Wang, Kota Yamaguchi, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2018**. Salt Lake City, Utah. June 2018.

Feedback Propagation Inference as a solution

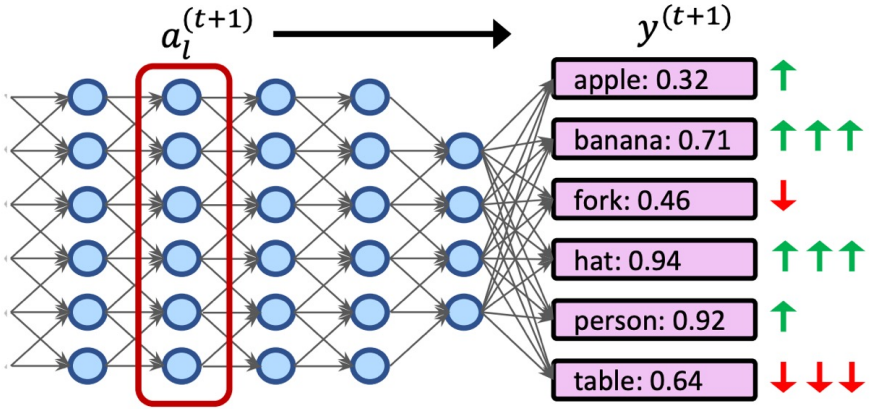


We found that output variables can be used as input at test time by iterative inference using backward and forward passes on intermediate features.

$$a_l^* = \text{argmin}_{a_l} L(Y_\text{k}, F_\text{k}^{(l)}(a_l, \Theta)),$$
$$\hat{Y}_\text{u} = F_\text{u}^{(l)}(a_l^*, \Theta).$$

Feedback-prop: Convolutional Neural Network Inference under Partial Evidence
Tianlu Wang, Kota Yamaguchi, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2018**. Salt Lake City, Utah. June 2018.

Feedback Propagation Inference as a solution

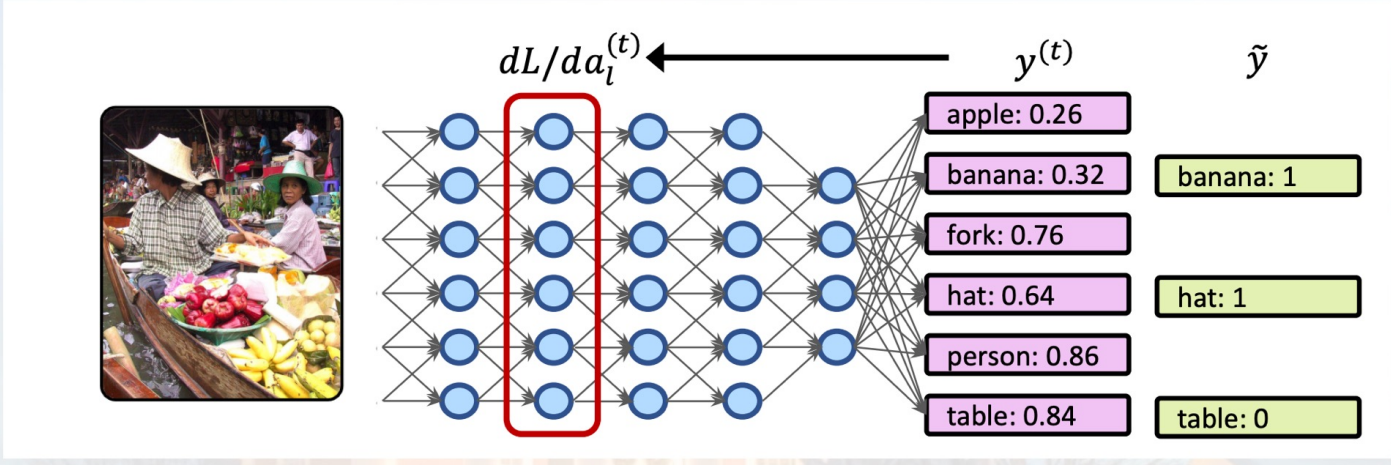


We found that output variables can be used as input at test time by iterative inference using backward and forward passes on intermediate features.

$$a_l^* = \text{argmin}_{a_l} L(Y_\text{k}, F_\text{k}^{(l)}(a_l, \Theta)),$$
$$\hat{Y}_\text{u} = F_\text{u}^{(l)}(a_l^*, \Theta).$$

Feedback-prop: Convolutional Neural Network Inference under Partial Evidence
Tianlu Wang, Kota Yamaguchi, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2018**. Salt Lake City, Utah. June 2018.

Feedback Propagation Inference as a solution

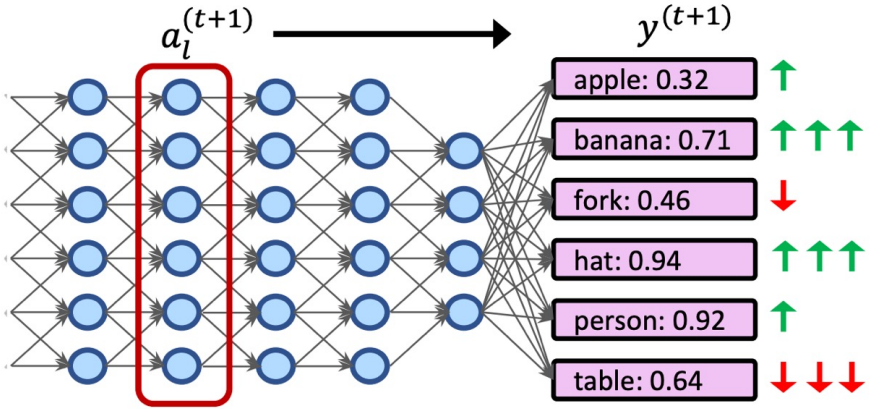


We found that output variables can be used as input at test time by iterative inference using backward and forward passes on intermediate features.

$$a_l^* = \text{argmin}_{a_l} L(Y_\text{k}, F_\text{k}^{(l)}(a_l, \Theta)),$$
$$\hat{Y}_\text{u} = F_\text{u}^{(l)}(a_l^*, \Theta).$$

Feedback-prop: Convolutional Neural Network Inference under Partial Evidence
Tianlu Wang, Kota Yamaguchi, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2018**. Salt Lake City, Utah. June 2018.

Feedback Propagation Inference as a solution



We found that output variables can be used as input at test time by iterative inference using backward and forward passes on intermediate features.

$$a_l^* = \text{argmin}_{a_l} L(Y_\text{k}, F_\text{k}^{(l)}(a_l, \Theta)),$$
$$\hat{Y}_\text{u} = F_\text{u}^{(l)}(a_l^*, \Theta).$$

Feedback-prop: Convolutional Neural Network Inference under Partial Evidence
Tianlu Wang, Kota Yamaguchi, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2018**. Salt Lake City, Utah. June 2018.

Image Captioning Models

Vision
Deep CNN

Recurrent
Text Decoder

Caption in German

Recurrent
Text Decoder

Caption in English

NEW! Using Visual Feature Space as a Pivot Across Languages

Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, Vicente Ordonez. Findings of Empirical Methods in Natural Language Processing. **Findings of EMNLP 2020**. Accepted September 2020. [pdf] [code] [bibtex]

At test time: (Image + German) to English

Vision
Deep CNN

Recurrent
Text Decoder

Caption in German

Recurrent
Text Decoder

Caption in English

NEW! Using Visual Feature Space as a Pivot Across Languages

Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, Vicente Ordonez. Findings of Empirical Methods in Natural Language Processing. **Findings of EMNLP 2020**. Accepted September 2020. [pdf] [code] [bibtex]

At test time: (Image + English) to German

Vision
Deep CNN

Recurrent
Text Decoder

Caption in German

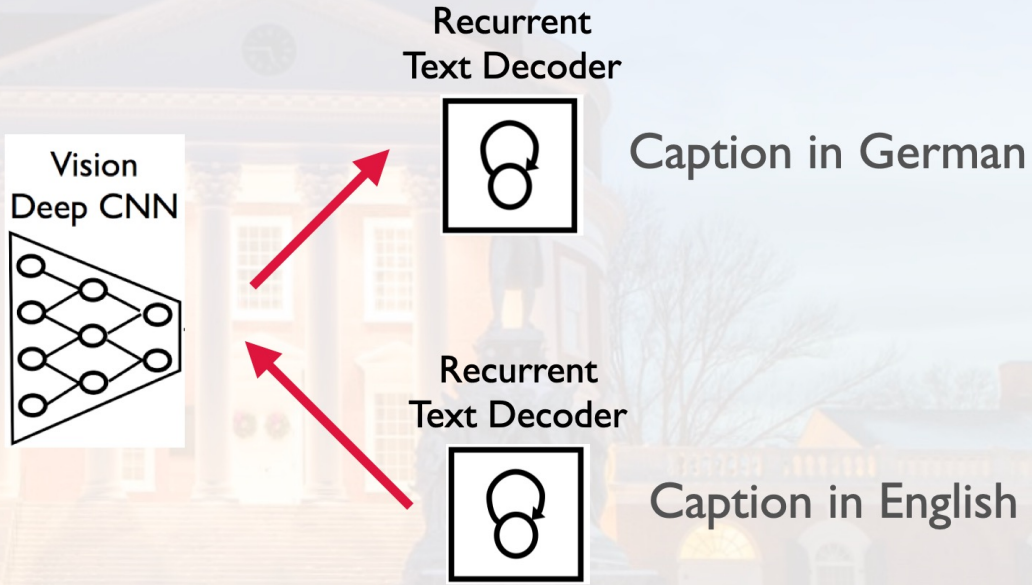
Recurrent
Text Decoder

Caption in English

NEW! Using Visual Feature Space as a Pivot Across Languages

Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, Vicente Ordonez. Findings of Empirical Methods in Natural Language Processing. **Findings of EMNLP 2020**. Accepted September 2020. [pdf] [code] [bibtex]

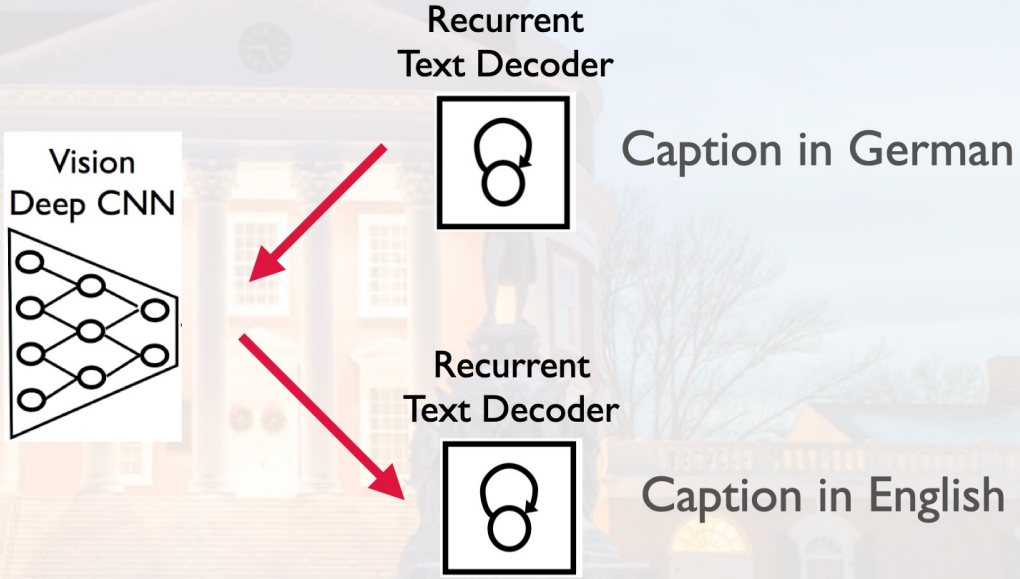
At test time: (English) to German



NEW! Using Visual Feature Space as a Pivot Across Languages

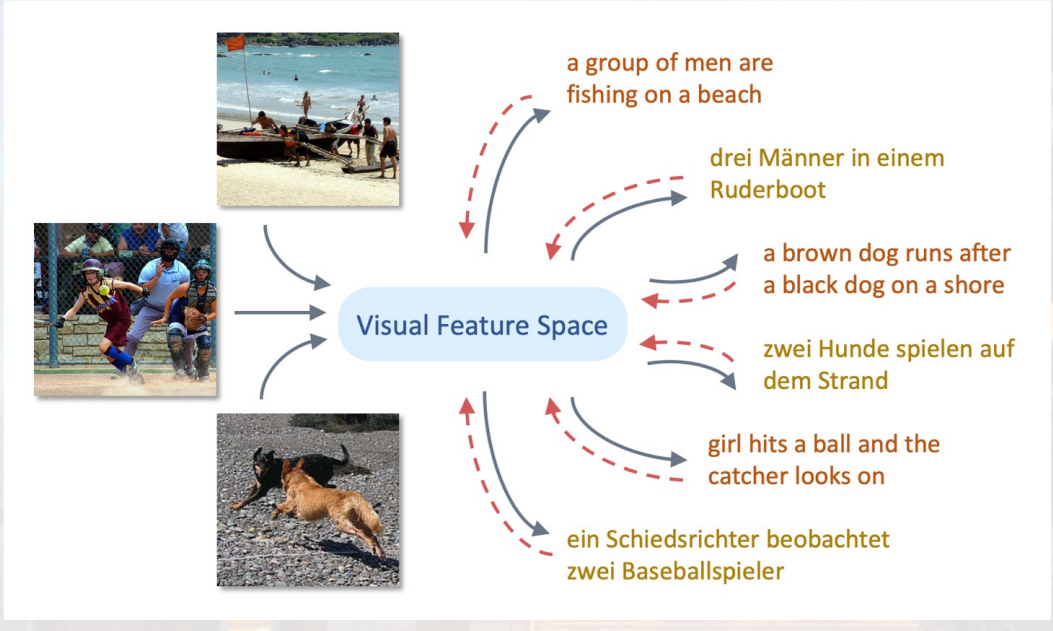
Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, Vicente Ordonez. Findings of Empirical Methods in Natural Language Processing. **Findings of EMNLP 2020**. Accepted September 2020. [pdf] [code] [bibtex]

At test time: (German) to English



NEW! Using Visual Feature Space as a Pivot Across Languages
Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, Vicente Ordonez. Findings of Empirical Methods in Natural Language Processing. **Findings of EMNLP 2020**. Accepted September 2020. [pdf] [code] [bibtex]

Using Visual Feature Space as a Pivot Across Languages



NEW! Using Visual Feature Space as a Pivot Across Languages

Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, Vicente Ordonez. Findings of Empirical Methods in Natural Language Processing. **Findings of EMNLP 2020**. Accepted September 2020. [pdf] [code] [bibtex]

Some Results

INPUTS

ein Mann fängt das Ball am Strand.

OUTPUTS

image: A man in a white shirt is jumping in the air.

image + de: A man is playing with a red ball on the beach.

Some Results

INPUTS

新聞紙 の 上 に 無数 の はさみ が おい て ある

OUTPUTS

image: A group of blue and white cake on a table.

image + jp: A table topped with lots of blue and white scissors.

Some Results

ein Kleinkind spielt mit einer gelben Plastikschippe.	a baby is playing with a yellow ball in the grass.
デスク の 上 に パソコン や ライト 、 本 が 置か れ て いる	a desk with a laptop and a book.

Questions?