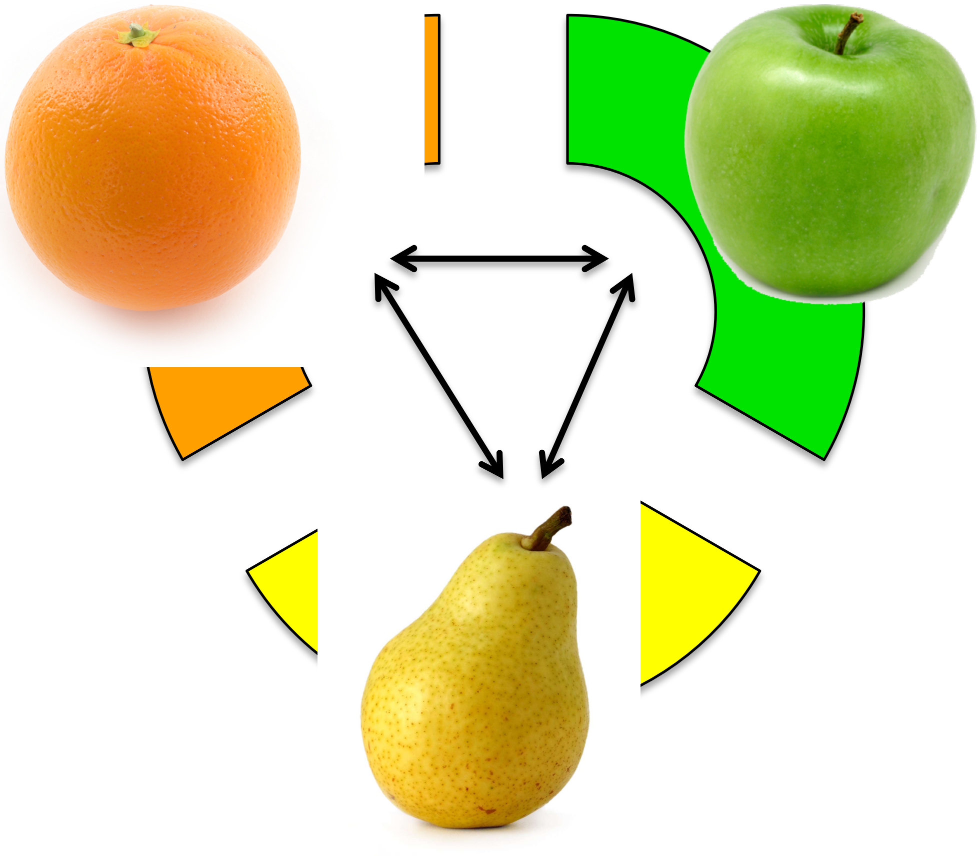
Karlsruhe Institute of Technology

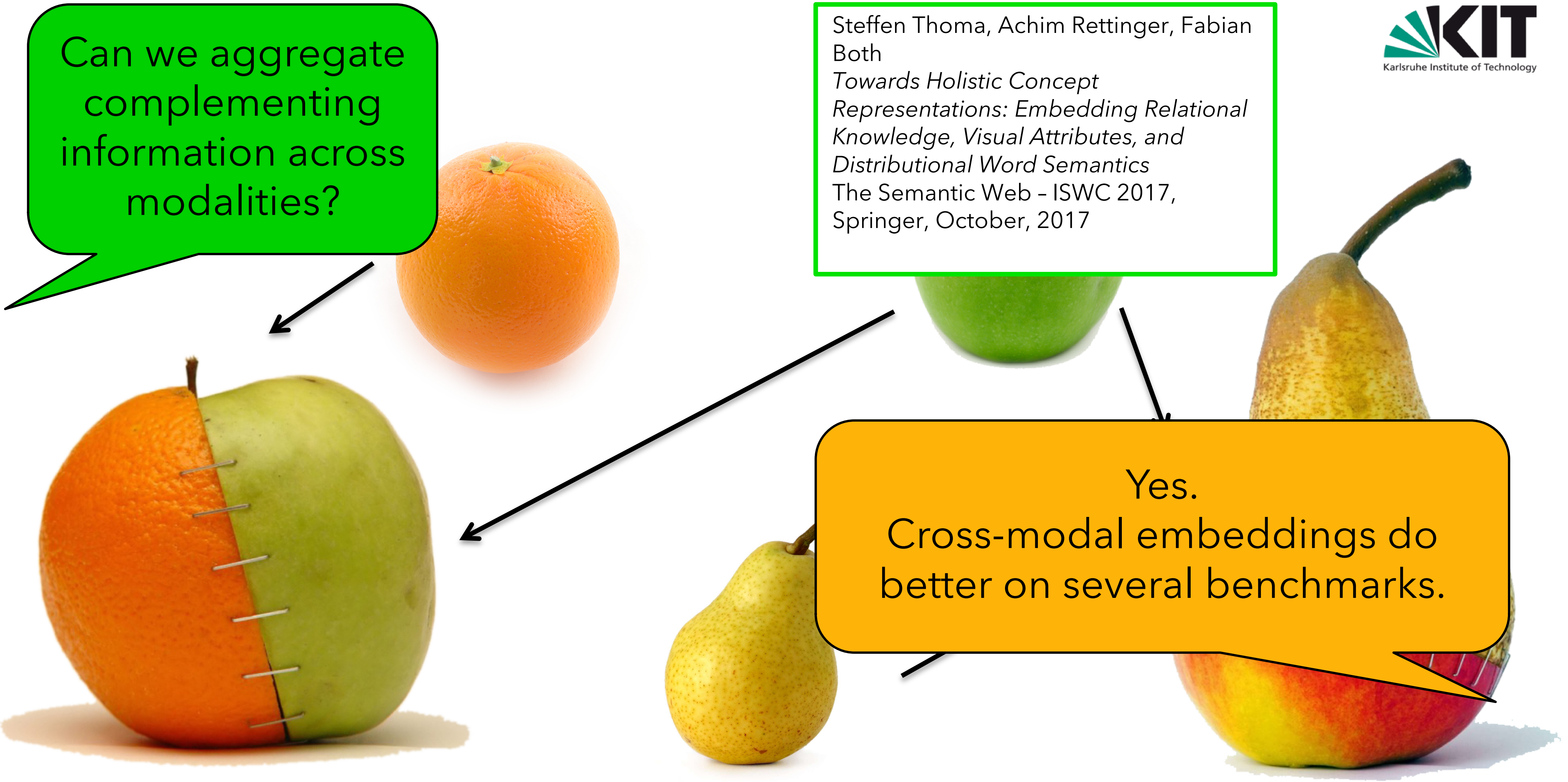
Knowledge Guided Attention and Inference for Describing Images Containing Unseen Objects

Aditya Mogadala, Umanga Bista, Lexing Xie, Achim Rettinger

rettinger@kit.edu, http://www.aifb.kit.edu/web/Achim_Rettinger/en, http://www.aifb.kit.edu/web/Inproceedings3603

ADAPTIVE DATA ANALYTICS GROUP
INSTITUTE OF APPLIED INFORMATICS AND FORMAL DESCRIPTION METHODS (AIFB)





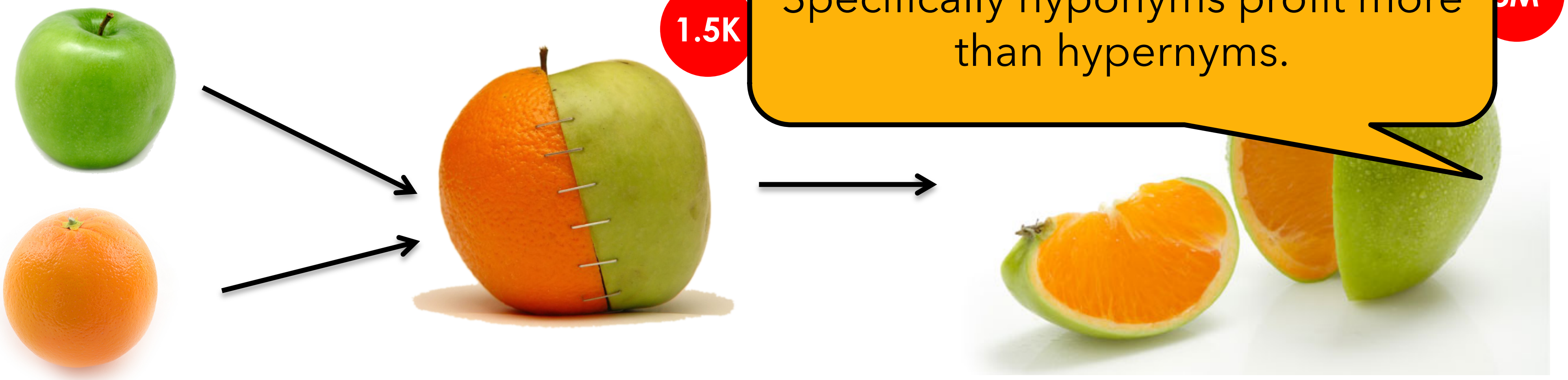
Can we aggregate complementing information across modalities?

Steffen Thoma, Achim Rettinger, Fabian Both
Towards Holistic Concept Representations: Embedding Relational Knowledge, Visual Attributes, and Distributional Word Semantics
The Semantic Web – ISWC 2017, Springer, October, 2017

Yes.
Cross-modal embeddings do better on several benchmarks.

Fabian Both, Steffen Thoma, Achim Rettinger.
Cross-modal Knowledge Transfer: Improving the Word Embedding of Apple by Looking at Oranges.
K-CAP2017, The 9th International Conference on Knowledge Capture, ACM, Dezember, 2017

Can we extrapolate cross-modal information to entities unseen in some of the other modalities?



Yes.
Specifically hyponyms profit more than hypernyms.

Aditya Mogadala, Umanga Bista, Lexing Xie and Achim Rettinger. *Knowledge Guided Attention and Inference for Describing Images Which Contain Unseen Objects,* ESWC 2018

Can we extrapolate knowledge about translating entities across modalities without having seen them during training?

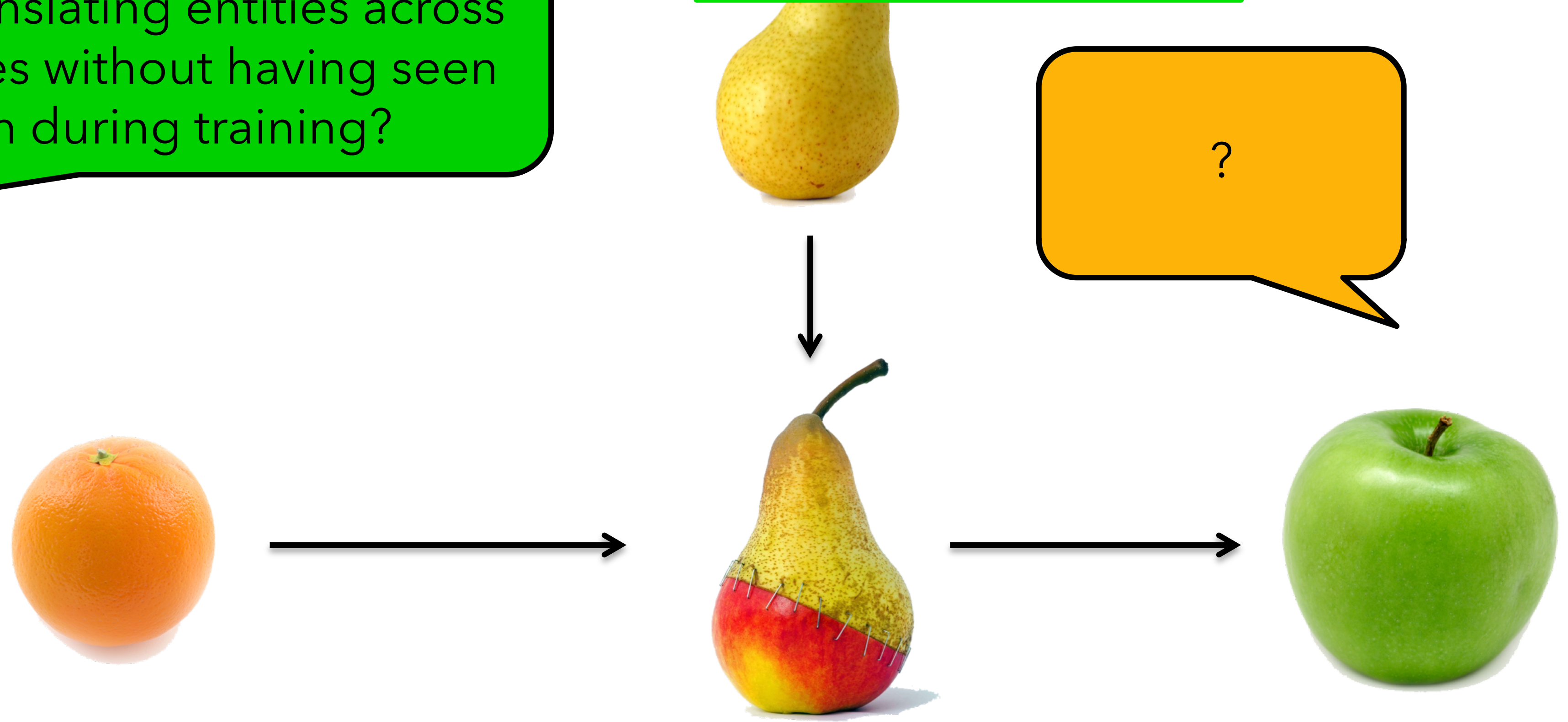


IMAGE CAPTION GENERATION

Visual Object Detection

Images on the Web depict a huge variety of visual objects

Truffle

Mammoth

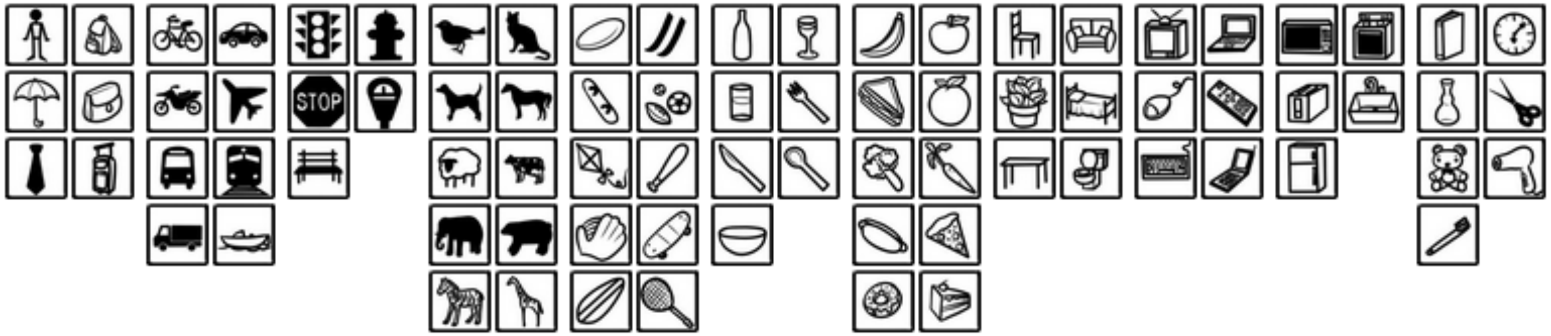
Blackbird

Papaya

642 Visual Object Categories by ImageNet

Description Generation for Images

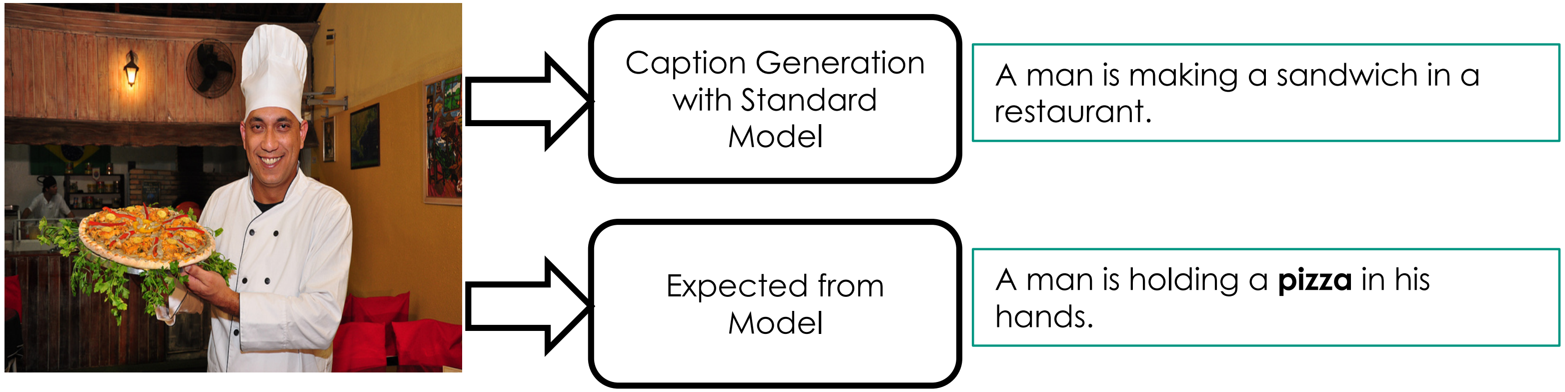
Training data for image captioning (i.e. image-caption pairs) cover only a fraction of objects that can be detected by image classifiers.



80 MSCOCO Visual Object Categories

Challenge - Missing Captions for Images

Parallel caption training examples are missing for images containing visual object category “**pizza**”.



Caption Generation with Standard Model

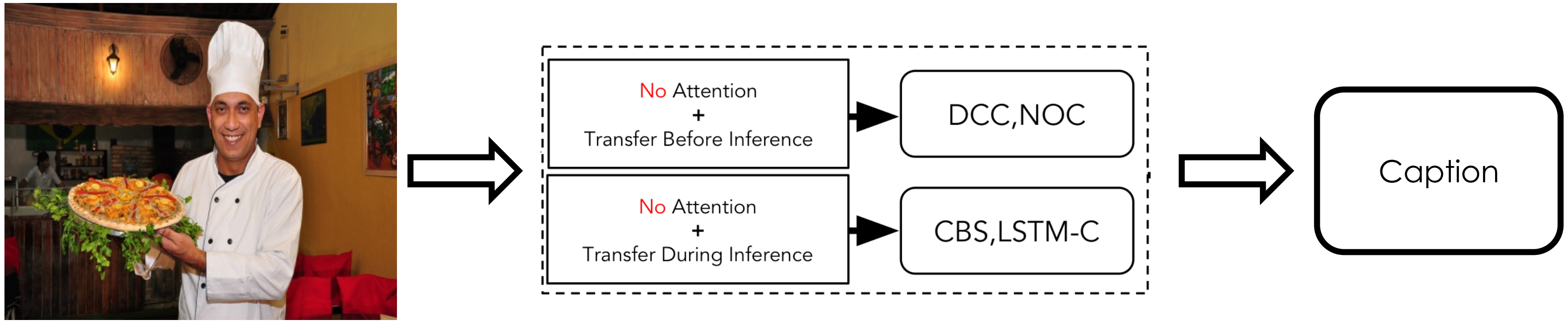
A man is making a sandwich in a restaurant.

Expected from Model

A man is holding a **pizza** in his hands.

Related Work

Approaches that can handle unseen objects.



No Attention
+
Transfer Before Inference

DCC,NOC

No Attention
+
Transfer During Inference

CBS,LSTM-C

Caption

Missing in Related Work

Attention

Our attention mechanism learns to focus on the salient aspects in the image for caption generation.

Inference

Transfer either before or during inference. We do both.

KNOWLEDGE GUIDED ATTENTION AND INFERENCE

Our Contributions

ESA

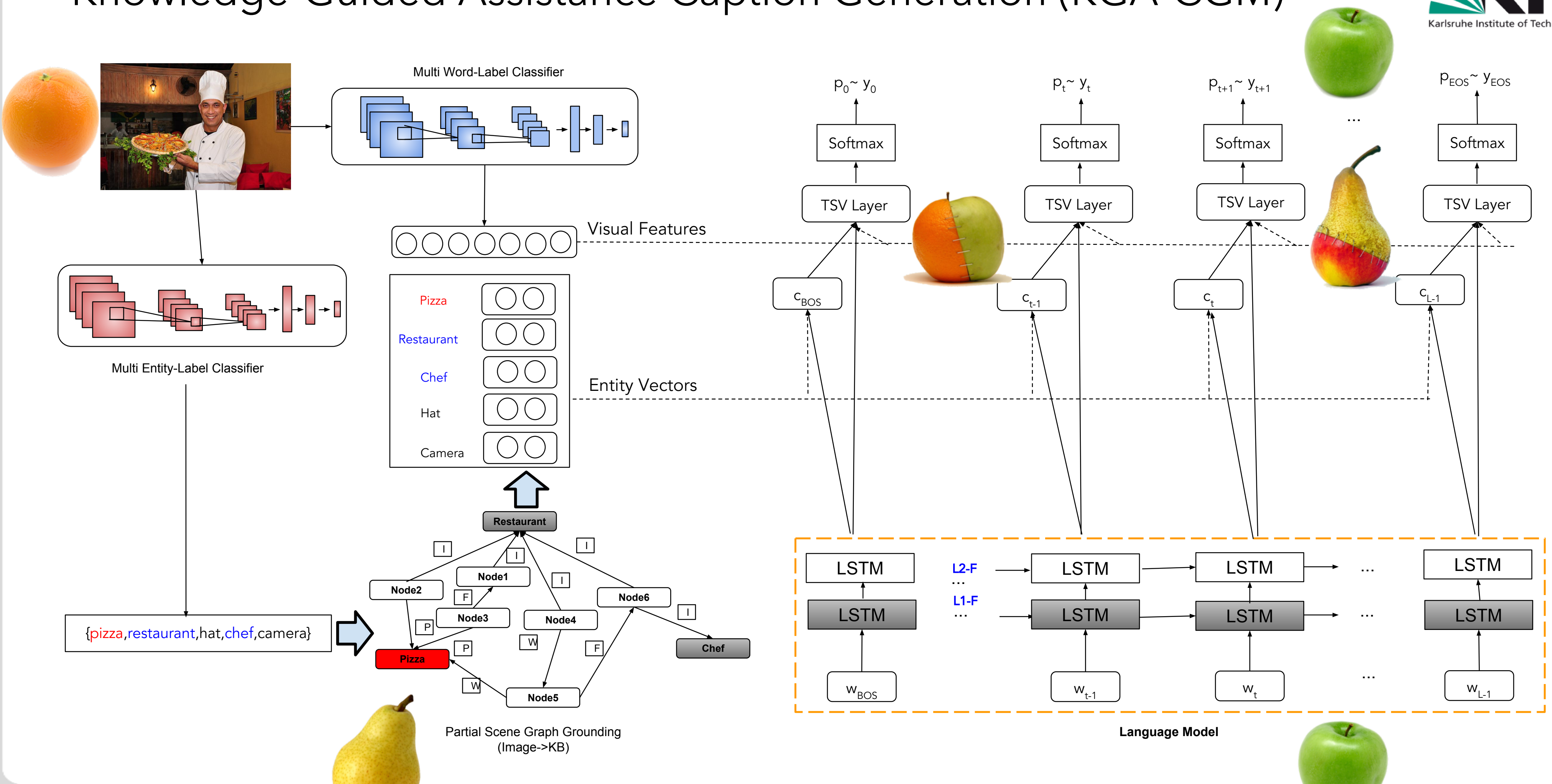
Introduce an attention mechanism into the caption generation model from External Semantic Knowledge (ESA) provided by a knowledge graph (KG)

CI

Constraint before and during Inference (CI) for transferring information between seen words and unseen visual object categories by exploiting external semantic knowledge provided by a knowledge graph (KG).

Knowledge-Guided Assistance Caption Generation (KGA-CGM)

Karlsruhe Institute of Technology

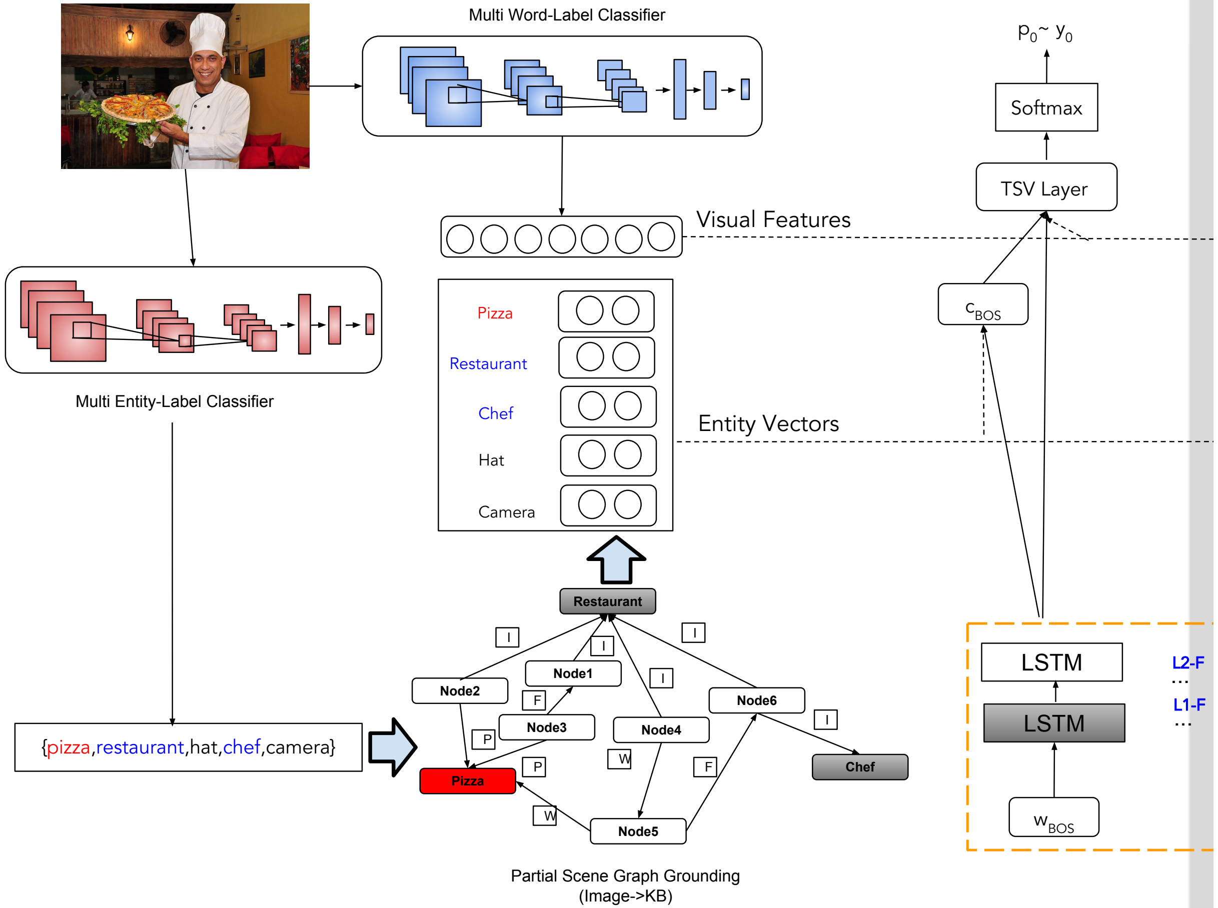


External Sematic Attention

$$\boldsymbol{\beta}_{ti} = \frac{exp(\boldsymbol{O}_{ti})}{\sum_{j=1}^{L} exp(\boldsymbol{O}_{tj})}$$

$$\boldsymbol{O}_{ti} = tanh((\boldsymbol{h}_t^2)^T W_{he} \boldsymbol{e}_i)$$

$$\boldsymbol{c}_t = \sum_{i=1}^{L} \beta_{ti} \boldsymbol{e}_i$$

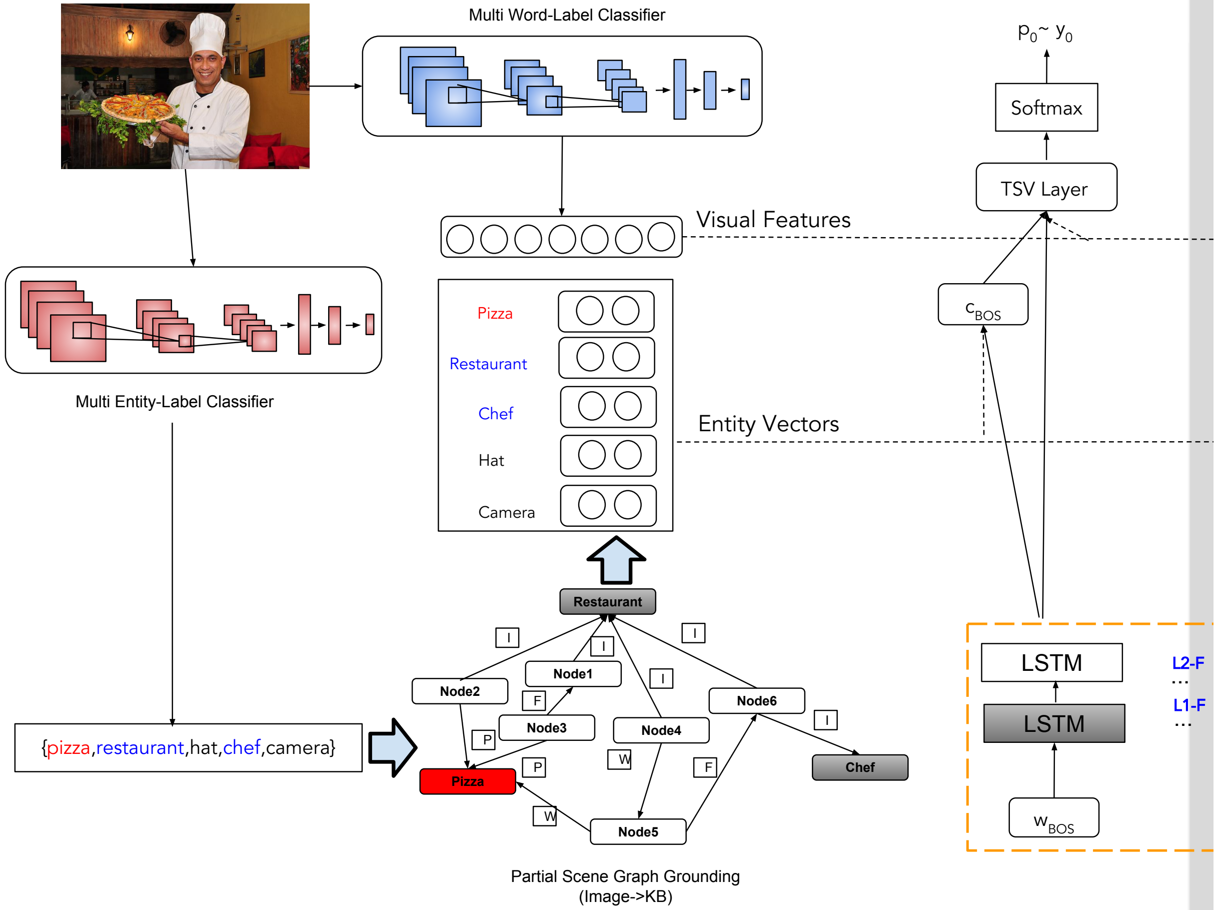


TSV Layer

$$\boldsymbol{TSV}_t = W_{\boldsymbol{h}_t^2}\boldsymbol{h}_t^2 + W_{\boldsymbol{c}_t}\boldsymbol{c}_t + W_{\boldsymbol{I}_t}\boldsymbol{I}_t$$

$$\boldsymbol{p}_{t+1} = softmax(\boldsymbol{TSV}_t)$$

$$\min_{\theta} -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{L^{(n)}} log(\boldsymbol{p}(y_t^{(n)}))$$



Inference – Generating unseen objects

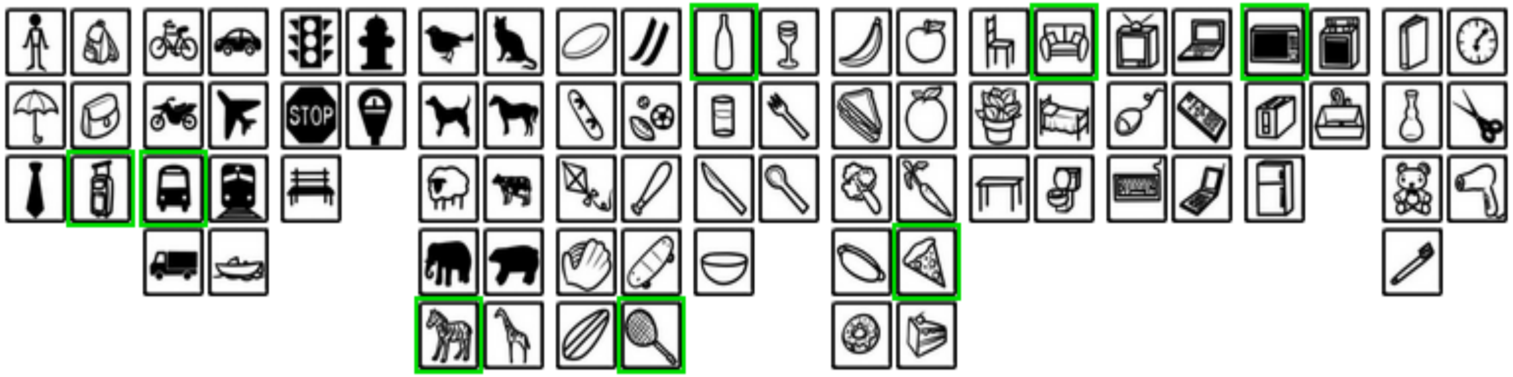
[UnseenObj17]

```
Input: M=$\{W_{he}, W_{\boldsymbol{h}_t^2}, W_{\boldsymbol{c}_t}, W_{\boldsymbol{I}_t}\}$
Output: $M_{new}$
Initialize List(closest) = cosine_distance(List(unseen),vocabulary) ;
Initialize $W_{\boldsymbol{c}_t}[v_{unseen},:]$, $W_{\boldsymbol{h}_t^2}[v_{unseen},:]$, $W_{\boldsymbol{I}_t}[v_{unseen},:] = 0$ ;
Function Before Inference
    forall items $T$ in closest and $Z$ in unseen do
        if $T$ and $Z$ is vocabulary then
            $W_{\boldsymbol{c}_t}[v_Z,:] = W_{\boldsymbol{c}_t}[v_T,:]$ ;
            $W_{\boldsymbol{h}_t^2}[v_Z,:] = W_{\boldsymbol{h}_t^2}[v_T,:]$ ;
            $W_{\boldsymbol{I}_t}[v_Z,:] = W_{\boldsymbol{I}_t}[v_T,:]$ ;
        end
        if $i_T$ and $i_Z$ in visual features then
            $W_{\boldsymbol{I}_t}[i_Z,i_T]=0$ ;
            $W_{\boldsymbol{I}_t}[i_T,i_Z]=0$ ;
        end
    end
    $M_{new}$ = M ;
    return $M_{new}$ ;
end
```

EVALUATION

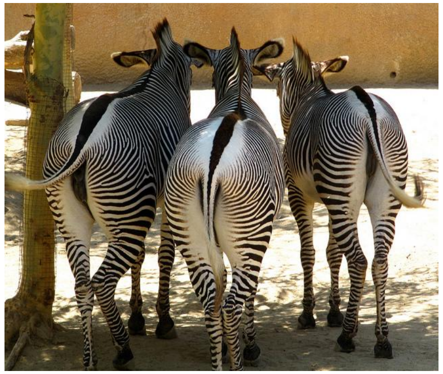
Evaluation Setup

- 8 held out objects from MSCOCO
- Image-Caption Pairs: 70K Training, 20K Validation, 20K Testing
- CNN Architectures: VGG16 [Simoyan et. Al. 2014]
- Unpaired Textual Corpus: British National Corpus, Wikipedia, SBU1M
- Entity Vectors: RDF2Vec [Ristoski et. Al. 2014]
- Evaluation Metrics: Meteor, Spice, F1



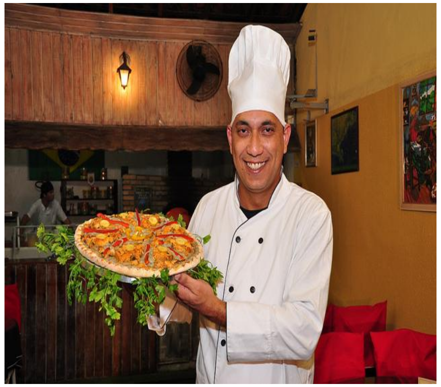
Microwave, Racket, Bottle, Zebra, Pizza, Couch , Bus, Suitcase

Qualitative Results



Unseen Object: Zebra

Predicted Entity-Labels (Top-3):Zebra,Enclosure,Zoo
Base: A couple of animals that are standing in a field
NOC: Zebras standing together in a field with zebras
KGA-CGM: A group of zebras standing in a line

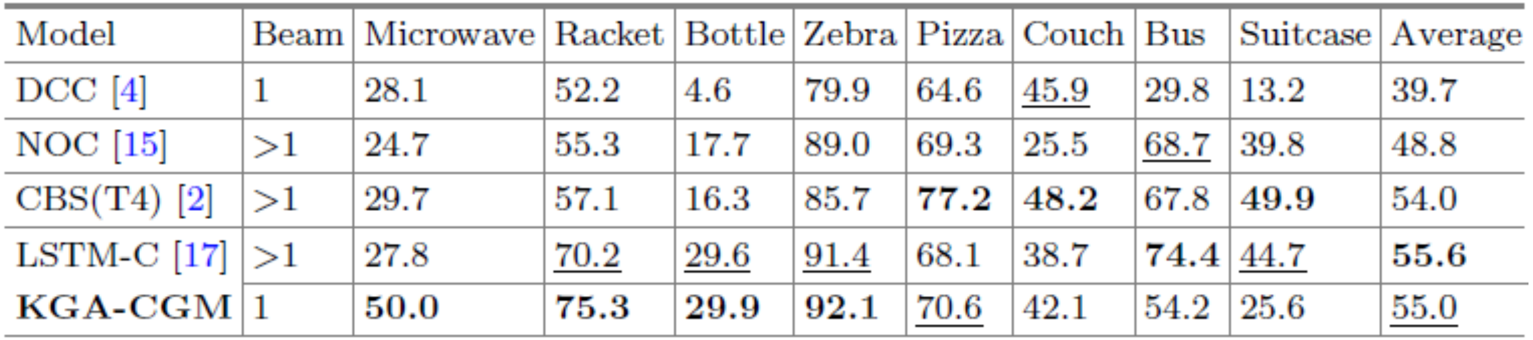


Unseen Object: Pizza

Predicted Entity-Labels (Top-3): Pizza,Restaurant,Hat
Base: A man is making a sandwich in a restaurant
NOC: A man standing next to a table with a pizza in front of it.
KGA-CGM: A man is holding a pizza in his hands

Quantitative Results

F1-Score

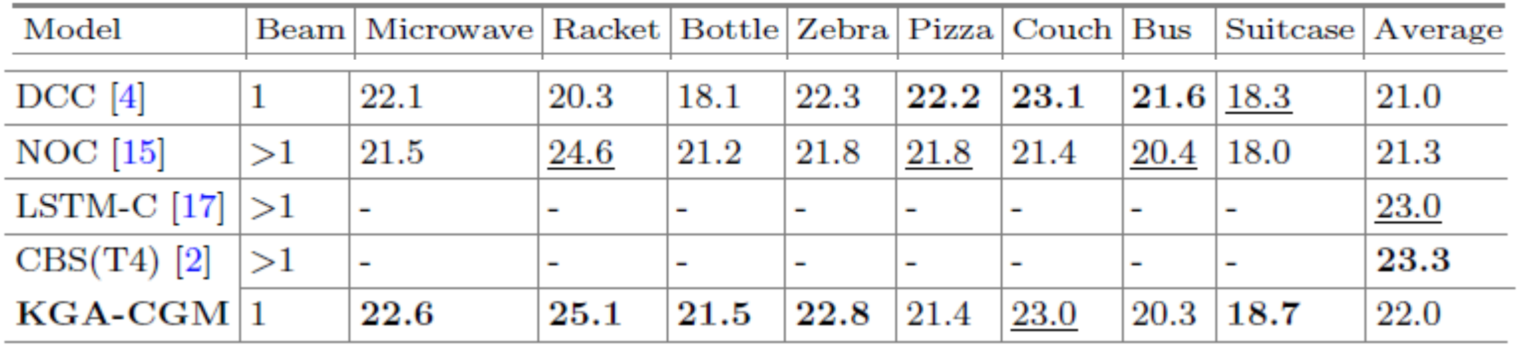


Model	Beam	Microwave	Racket	Bottle	Zebra	Pizza	Couch	Bus	Suitcase	Average
DCC [4]	1	28.1	52.2	4.6	79.9	64.6	45.9	29.8	13.2	39.7
NOC [15]	>1	24.7	55.3	17.7	89.0	69.3	25.5	68.7	39.8	48.8
CBS(T4) [2]	>1	29.7	57.1	16.3	85.7	**77.2**	**48.2**	67.8	**49.9**	54.0
LSTM-C [17]	>1	27.8	70.2	29.6	91.4	68.1	38.7	**74.4**	44.7	**55.6**
KGA-CGM	1	**50.0**	**75.3**	**29.9**	**92.1**	70.6	42.1	54.2	25.6	55.0

KGA-CGM (our proposed model). Underline represent second best

Quantitative Results

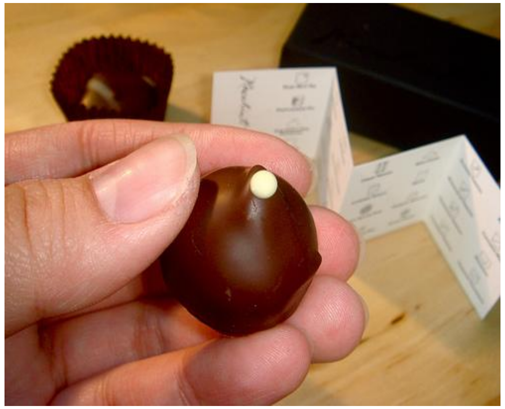
METEOR



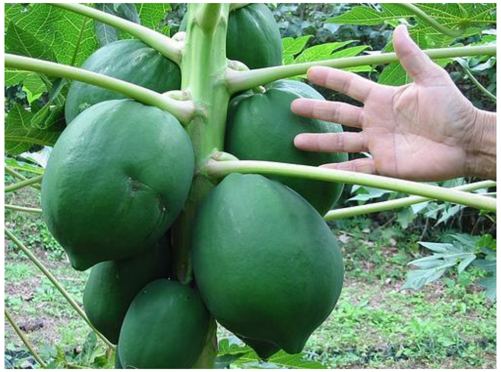
Model	Beam	Microwave	Racket	Bottle	Zebra	Pizza	Couch	Bus	Suitcase	Average
DCC [4]	1	22.1	20.3	18.1	22.3	**22.2**	**23.1**	**21.6**	18.3	21.0
NOC [15]	>1	21.5	24.6	21.2	21.8	21.8	21.4	20.4	18.0	21.3
LSTM-C [17]	>1	-	-	-	-	-	-	-	-	23.0
CBS(T4) [2]	>1	-	-	-	-	-	-	-	-	**23.3**
KGA-CGM	1	**22.6**	**25.1**	**21.5**	**22.8**	21.4	23.0	20.3	**18.7**	22.0

KGA-CGM (our proposed model) and underline represent second best

Scaling it by an order of magnitude

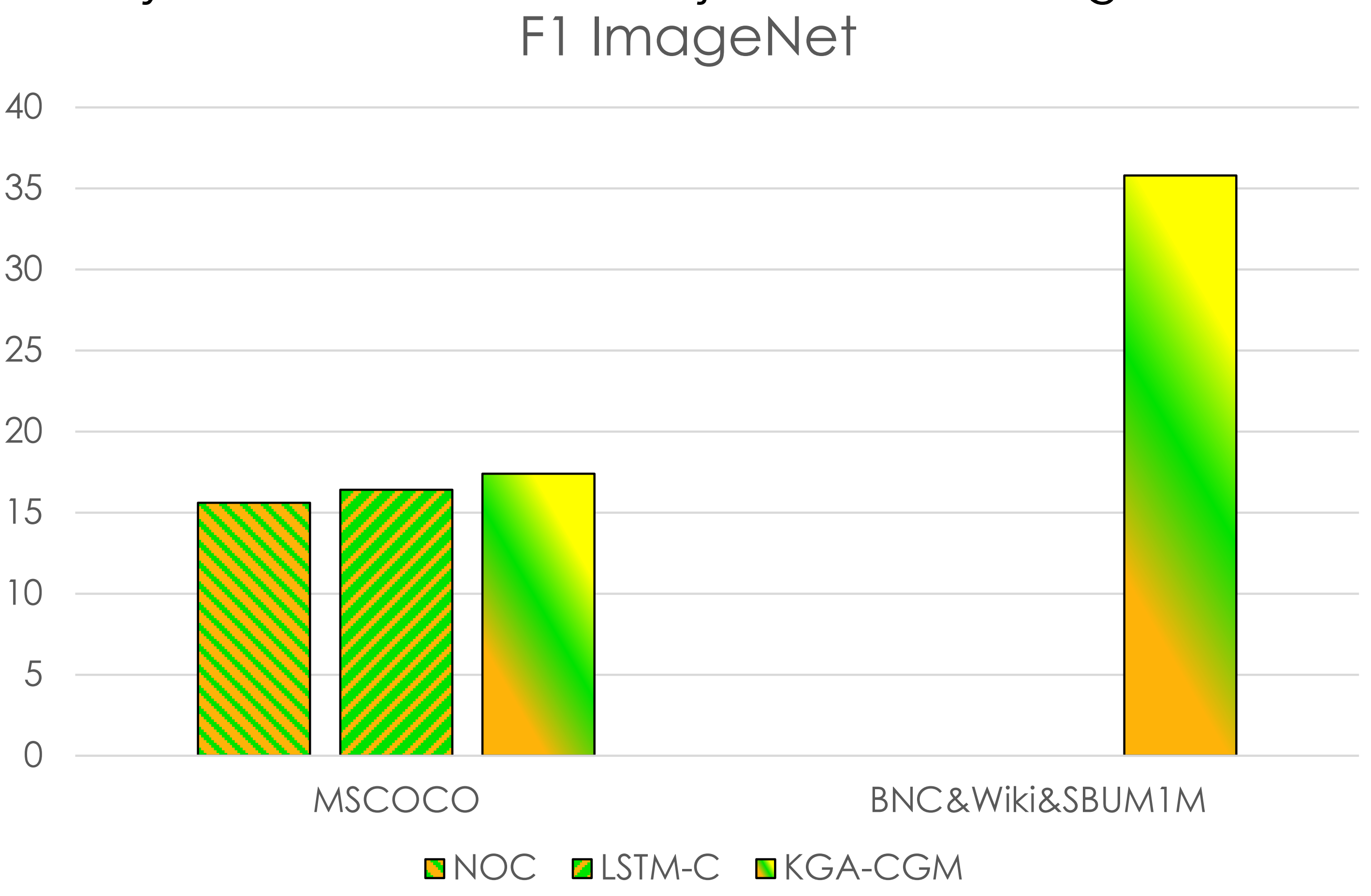


Unseen Object: Truffle
Guidance Before Inference: food → truffle
Base: A person holding a piece of paper.
KGA-CGM: A close up of a person holding truffle

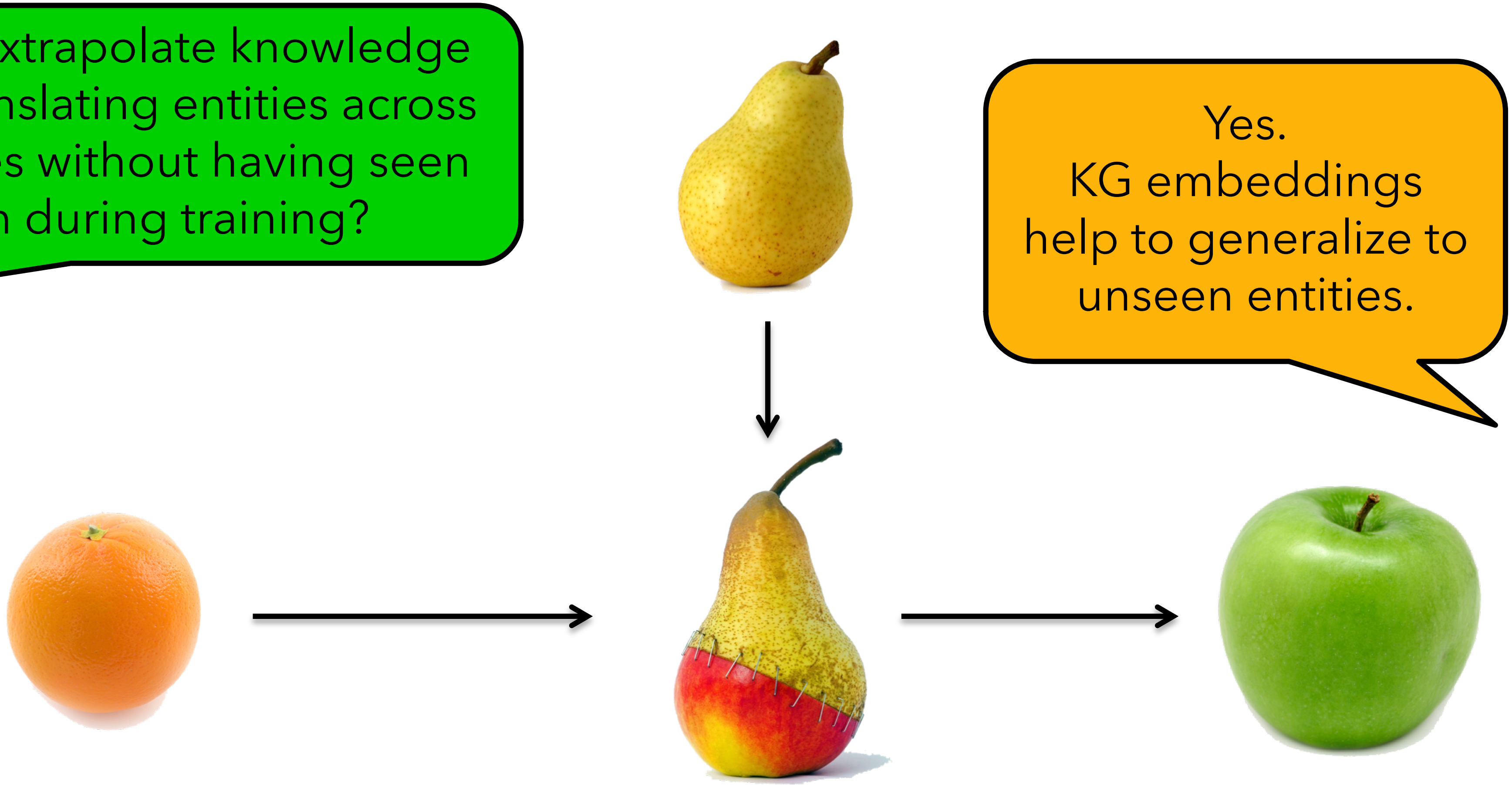


Unseen Object: Papaya
Guidance Before Inference: banana → papaya
Base: A woman standing in a garden.
KGA-CGM: These are ripe papaya hanging on a tree

Quantitative Analyse: Out-of-domain Objektbeschreibung



Can we extrapolate knowledge about translating entities across modalities without having seen them during training?



Yes.
KG embeddings help to generalize to unseen entities.

References

[Hendricks et al. 2016] Anne Hendricks, L., Venugopalan, S., Rohrbach, M., Mooney, R.,Saenko, K. and Darrell, T., 2016. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-10).

[Venugopalan et al. 2017] Venugopalan, S., Hendricks, L.A., Rohrbach, M., Mooney, R., Darrell, T. and Saenko, K., 2017. Captioning images with diverse objects. CVPR.

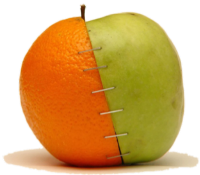
[Anderson et al. 2017] Anderson, P., Fernando, B., Johnson, M. and Gould, S., 2017. Guided Open Vocabulary Image Captioning with Constrained Beam Search. EMNLP.

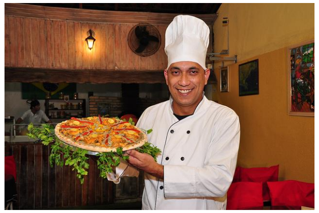
[Yao et al. 2017] Yao, T., Pan, Y., Li, Y. and Mei, T., 2017. Incorporating copying mechanism in image captioning for learning novel objects. CVPR.

[Simoyan et al. 2014] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[Ordonez et al. 2014] Ordonez, V., Kulkarni, G. and Berg, T.L., 2011. Im2text: Describing images using 1 million captioned photographs. In Advances in Neural Information Processing Systems (pp. 1143-1151).

[Ristoski et al. 2014] Ristoski, P. and Paulheim, H., 2016, October. Rdf2vec: Rdf graph embeddings for data mining. In International Semantic Web Conference (pp. 498-514). Springer International Publishing.





A man is holding a **pizza** in his hands.

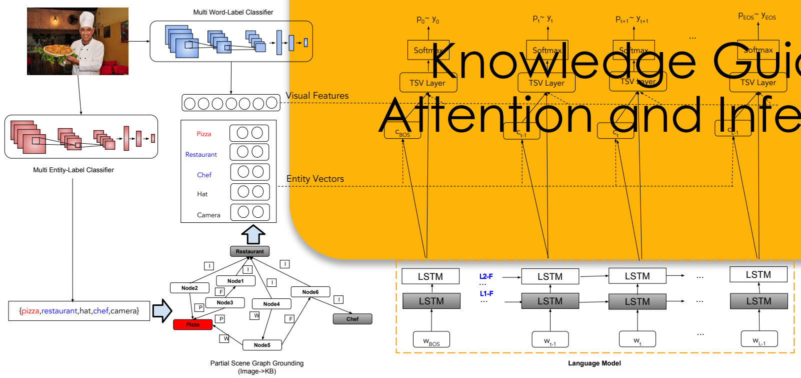
Karlsruhe Institute of Technology

Crossmodal Representation Learning and Transfer

Image Caption Generation

rettinger@kit.edu
http://www.aifb.kit.edu/web/Inproceedings3603

Knowledge Guided Attention and Inference



Evaluation

Unseen Object: Truffle
Guidance Before Inference: food → truffle
Base: A person holding a piece of paper.
KGA-CGM: A close up of a person holding truffle