# *A young boy with a tree made of trees:* Domain Adaptation of an LSTM-based Image Caption Generation Model

A project by Dominik Künkele and Maria Irena Szawerna for the LT2318 HT22 Artificial Intelligence: Cognitive Systems course.

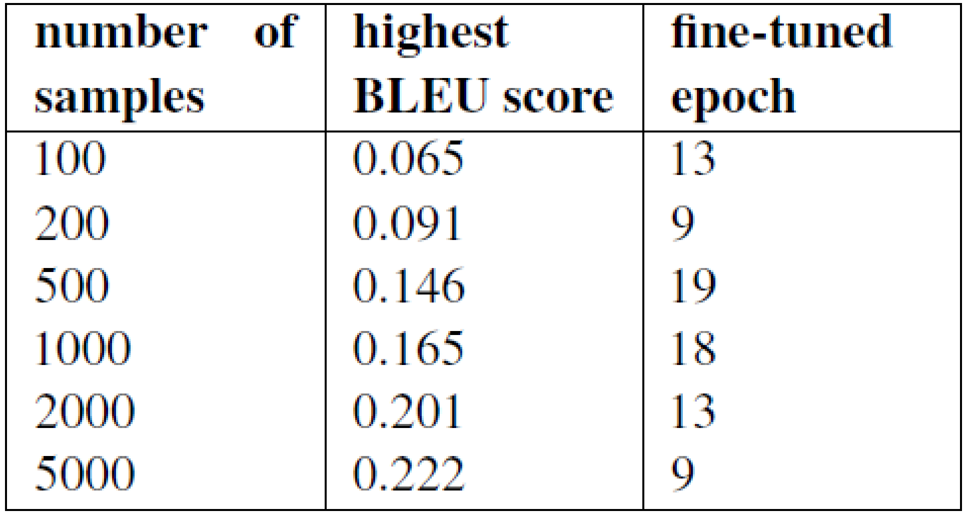
# Introduction (Maria)

1. Usefulness of image captioning software.
2. Do we need to re-train models for every domain?
3. CNN+LSTM image captioning models.
4. Can we adapt a pretrained model to a new domain?

# Materials and Methods (Dominik)

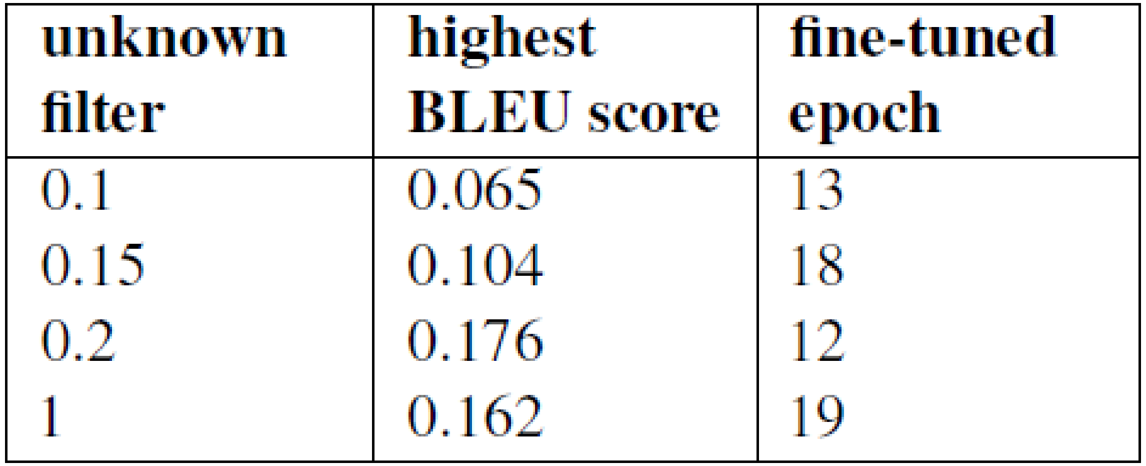
1. The image captioning model from Xu et al. 2015, implemented by Nikolai for one of the tutorials in the course.
2. Changes:
    1. Relation filter(s)
    2. Unknown filter
    3. Excluding UNK tokens from caption generation.
3. Flickr8k (original) and imageCLEF (new domain)
4. Human judgement questionnaire

# Results (Dominik)



| number of samples | highest BLEU score | fine-tuned epoch |
|---|---|---|
| 100 | 0.065 | 13 |
| 200 | 0.091 | 9 |
| 500 | 0.146 | 19 |
| 1000 | 0.165 | 18 |
| 2000 | 0.201 | 13 |
| 5000 | 0.222 | 9 |

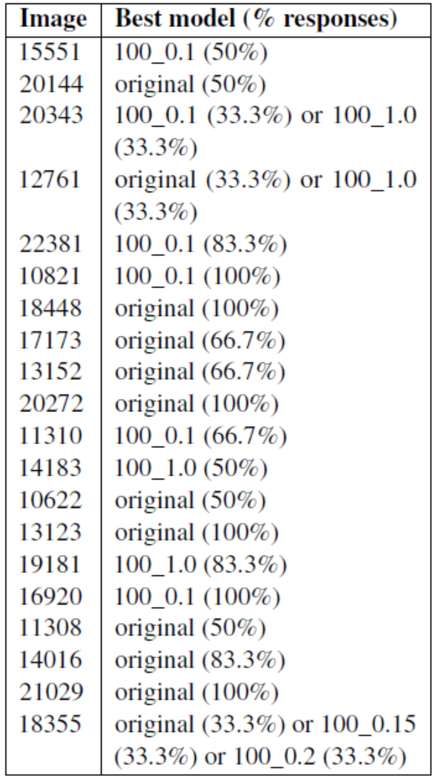
Table 1: BLEU scores for different sample sizes (unknown filter 0.1)



| unknown filter | highest BLEU score | fine-tuned epoch |
|---|---|---|
| 0.1 | 0.065 | 13 |
| 0.15 | 0.104 | 18 |
| 0.2 | 0.176 | 12 |
| 1 | 0.162 | 19 |

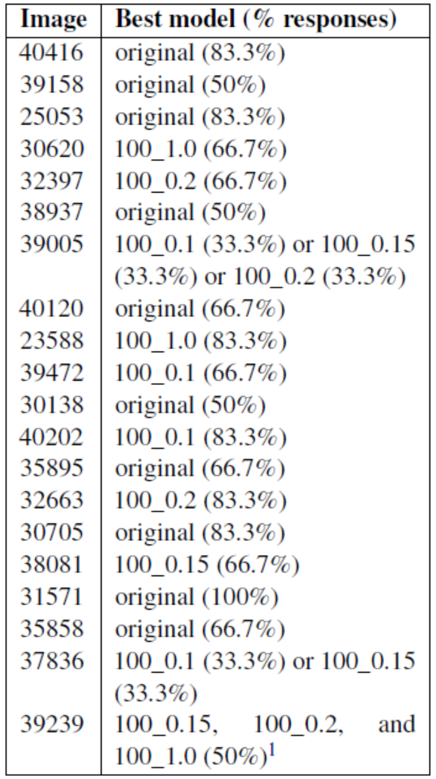
Table 2: BLEU scores for different unknown filters (number of samples: 100)

# Results (Maria)



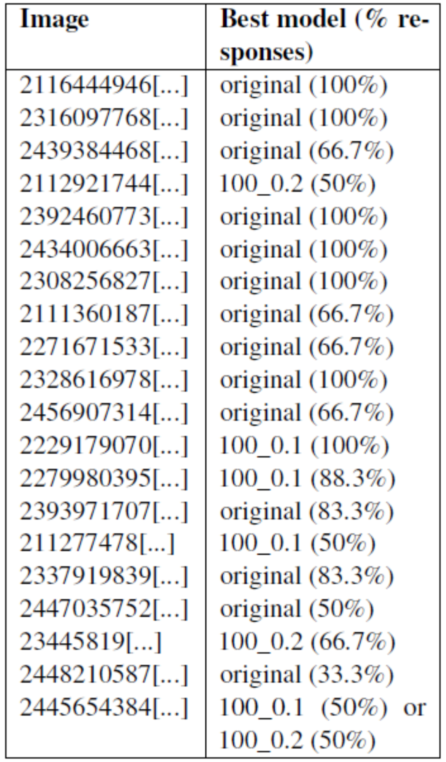
| Image | Best model (% responses) |
|---|---|
| 15551 | 100_0.1 (50%) |
| 20144 | original (50%) |
| 20343 | 100_0.1 (33.3%) or 100_1.0 (33.3%) |
| 12761 | original (33.3%) or 100_1.0 (33.3%) |
| 22381 | 100_0.1 (83.3%) |
| 10821 | 100_0.1 (100%) |
| 18448 | original (100%) |
| 17173 | original (66.7%) |
| 13152 | original (66.7%) |
| 20272 | original (100%) |
| 11310 | 100_0.1 (66.7%) |
| 14183 | 100_1.0 (50%) |
| 10622 | original (50%) |
| 13123 | original (100%) |
| 19181 | 100_1.0 (83.3%) |
| 16920 | 100_0.1 (100%) |
| 11308 | original (50%) |
| 14016 | original (83.3%) |
| 21029 | original (100%) |
| 18355 | original (33.3%) or 100_0.15 (33.3%) or 100_0.2 (33.3%) |

Table 3: Best captions per image in the unfiltered imageCLEF test set.



| Image | Best model (% responses) |
|---|---|
| 40416 | original (83.3%) |
| 39158 | original (50%) |
| 25053 | original (83.3%) |
| 30620 | 100_1.0 (66.7%) |
| 32397 | 100_0.2 (66.7%) |
| 38937 | original (50%) |
| 39005 | 100_0.1 (33.3%) or 100_0.15 (33.3%) or 100_0.2 (33.3%) |
| 40120 | original (66.7%) |
| 23588 | 100_1.0 (83.3%) |
| 39472 | 100_0.1 (66.7%) |
| 30138 | original (50%) |
| 40202 | 100_0.1 (83.3%) |
| 35895 | original (66.7%) |
| 32663 | 100_0.2 (83.3%) |
| 30705 | original (83.3%) |
| 38081 | 100_0.15 (66.7%) |
| 31571 | original (100%) |
| 35858 | original (66.7%) |
| 37836 | 100_0.1 (33.3%) or 100_0.15 (33.3%) |
| 39239 | 100_0.15, 100_0.2, and 100_1.0 (50%)[1] |

Table 4: Best captions per image in the filtered imageCLEF test set.



| Image | Best model (% responses) |
|---|---|
| 2116444946[...] | original (100%) |
| 2316097768[...] | original (100%) |
| 2439384468[...] | original (66.7%) |
| 2112921744[...] | 100_0.2 (50%) |
| 2392460773[...] | original (100%) |
| 2434006663[...] | original (100%) |
| 2308256827[...] | original (100%) |
| 2111360187[...] | original (66.7%) |
| 2271671533[...] | original (66.7%) |
| 2328616978[...] | original (100%) |
| 2456907314[...] | original (66.7%) |
| 2229179070[...] | 100_0.1 (100%) |
| 2279980395[...] | 100_0.1 (88.3%) |
| 2393971707[...] | original (83.3%) |
| 211277478[...] | 100_0.1 (50%) |
| 2337919839[...] | original (83.3%) |
| 2447035752[...] | original (50%) |
| 23445819[...] | 100_0.2 (66.7%) |
| 2448210587[...] | original (33.3%) |
| 2445654384[...] | 100_0.1 (50%) or 100_0.2 (50%) |

Table 5: Best captions per image in the Flickr8k test set.

Overall: captions were deemed to be low-quality, *rubbish*, to quote one of the participants.

# Discussion (Dominik)

1. The fine-tuning was much more disruptive than expected.
2. Differences in caption structure (syntax) and vocabulary are major.
    1. Less data – grammatical but thematically unfitting captions.
    2. More data – ungrammatical, theoretically thematically fitting captions (UNK tokens).
3. Small improvements based on human judgements, but the captions are still bad.

# Conclusions (Maria)

1. We have addressed our questions and found out what issues impede domain adapatation.
2. Ideas for future research:
    1. Testing the influence of other hyperparameters.
    2. Testing the same thing on a different model architecture.
    3. Fine-tuning on a more similar dataset.
    4. Exploring ways of mitigating the discovered issues.

# Thank you for your attention!