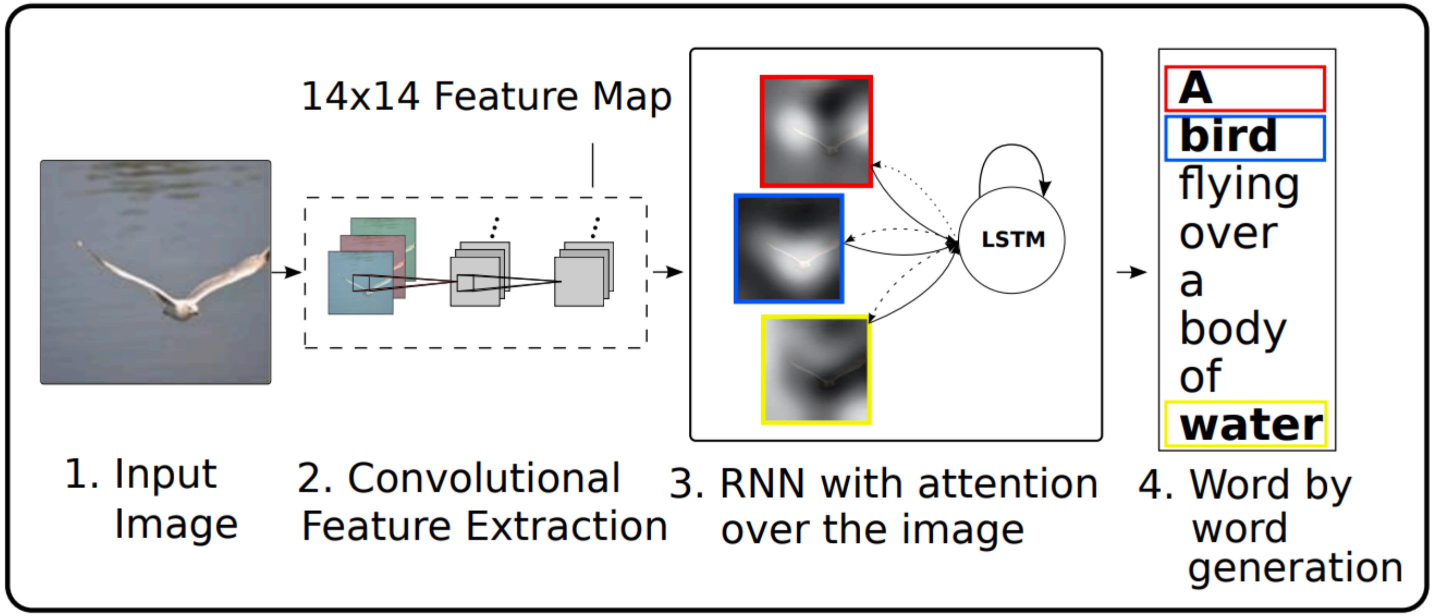
Neural Image Captioning with Attention

Implementierung des Papers:
“Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”

Neural Image Captioning

- Ziel: Implementierung von Attention in Image Captioning
- Grundlage: Paper " Show, Attend and Tell: Neural Image Caption Generation with Visual Attention"
- Modell: Encoder (VGG19) und Decoder (LSTM)
- Daten: Flickr8k

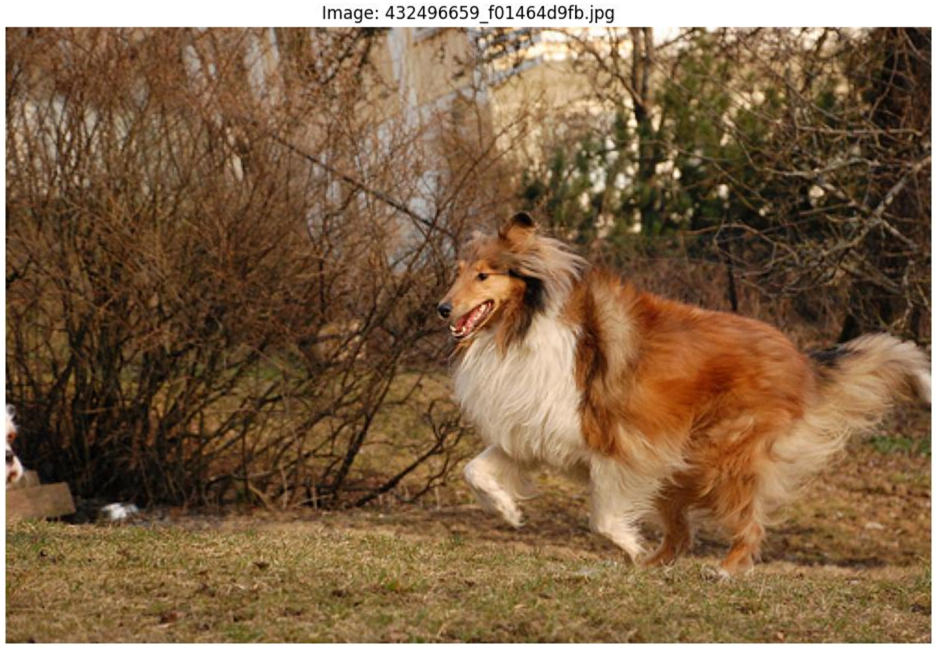

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Vokabular

- Wörter für Computer verarbeitbar machen: Tokens
- Kein BPE
- Spezialtokens
  - <sos>
  - <eos>
  - <pad> => collate Batches
  - <unk> => für Wörter mit weniger als 3 Vorkommnissen

Dataset

- Flickr8k
- Transformationen wie ImageNet_1k
- Padding der Sequenzen
- Gruppieren nach Bilder für Split

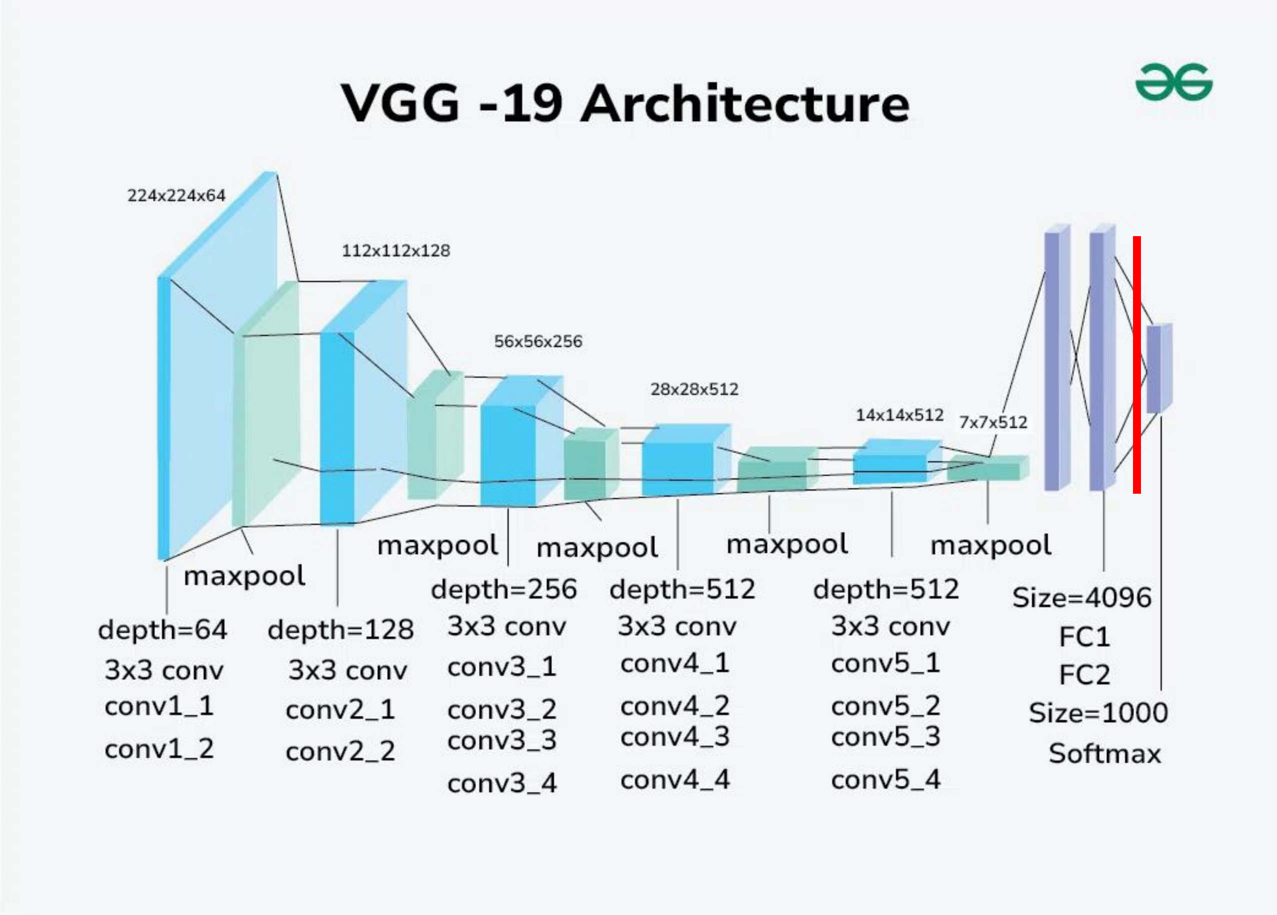


Captions:

1. a brown and white dog is walking through some wasteland .
2. A collie runs across a yard in the springtime .
3. A Collie surrounded by leafless bushes .
4. A dog that looks like Lassie walking in the fields .
5. The dog is in front of some bushes .

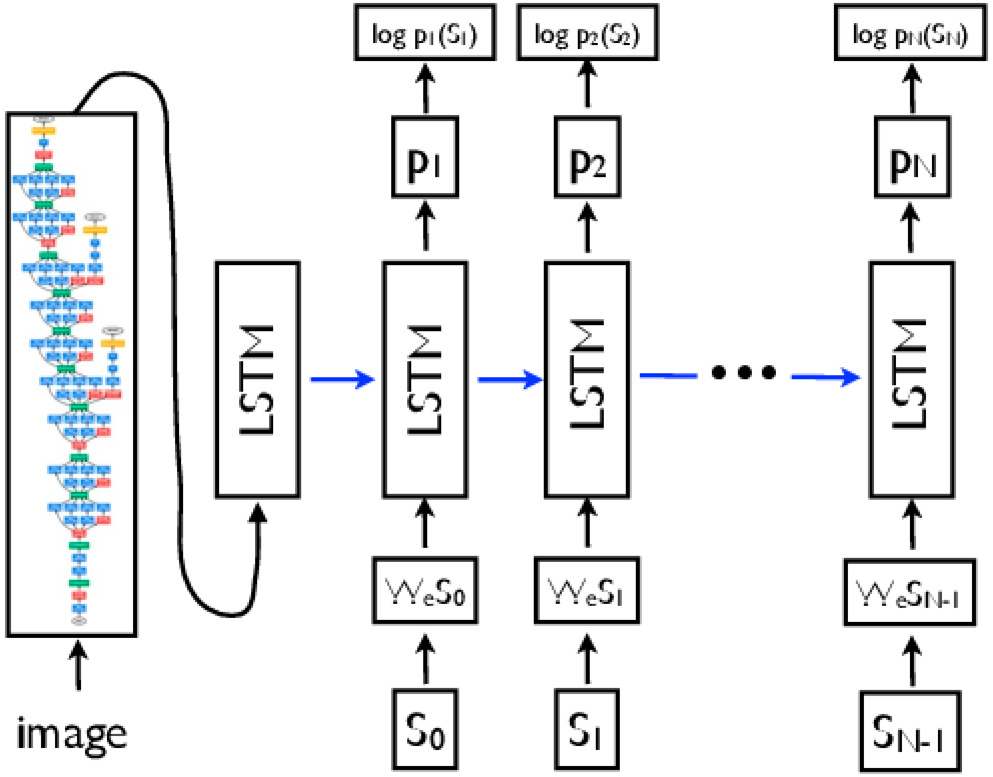
Modell Show and Tell - Encoder

- Encoder: VGG19
- Ohne letzten (Classification-)Layer
- Output=init State von LSTM



Modell Show and Tell - Decoder

- Embedding-Layer für bisherige Tokens
- LSTM
- Classification Layer für nächstes Token



Modell Show, **Attend** and Tell - Attention

- Inspiriert vom menschlichen Sehen
- Soft-Attention und Cross-Attention
- Regularisierung auf sum()=1
- Attention Gating

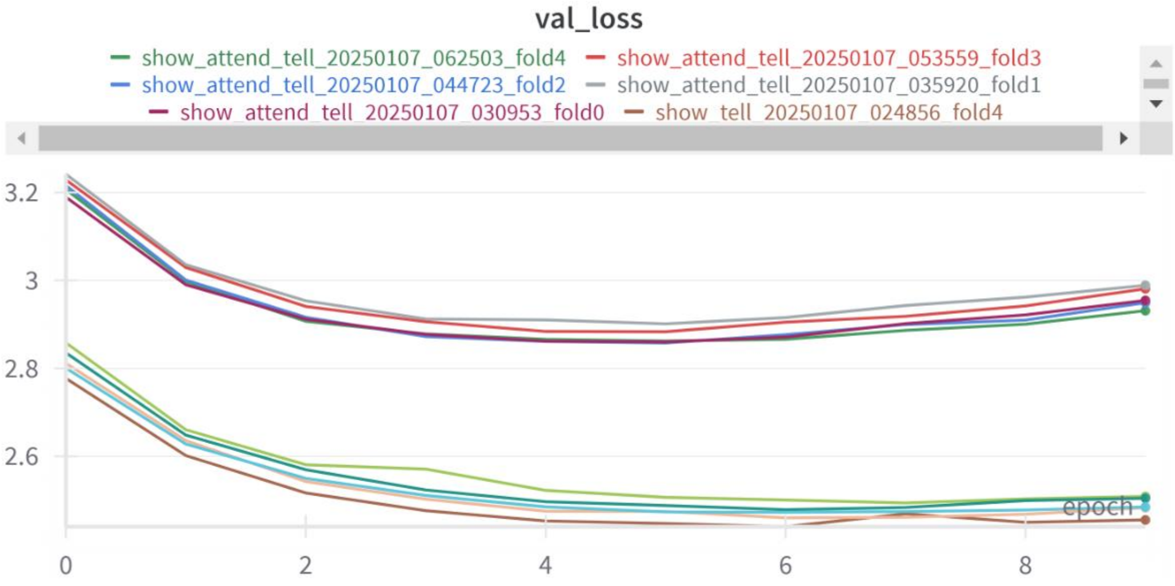
$$score = v^T tanh(W_1 h_i + W_2 s_t)$$

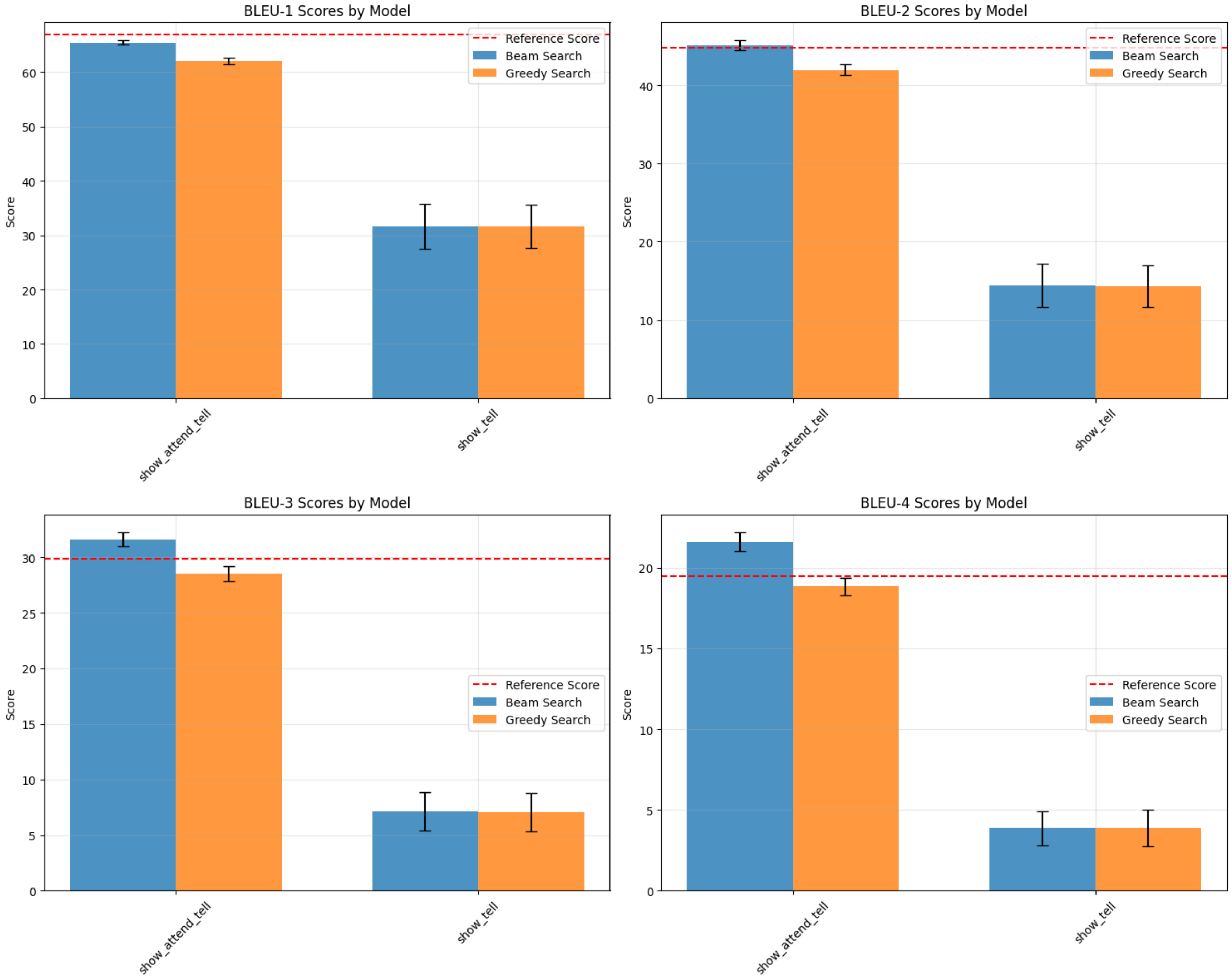
$$\alpha = softmax(score)$$

$$context = \Sigma \alpha_i h_i$$

Trainingsprozess

- Early Stopping auf Validation-Loss
- Teacher Forcing
- Bereits nach 6-8 Epochen overfitting





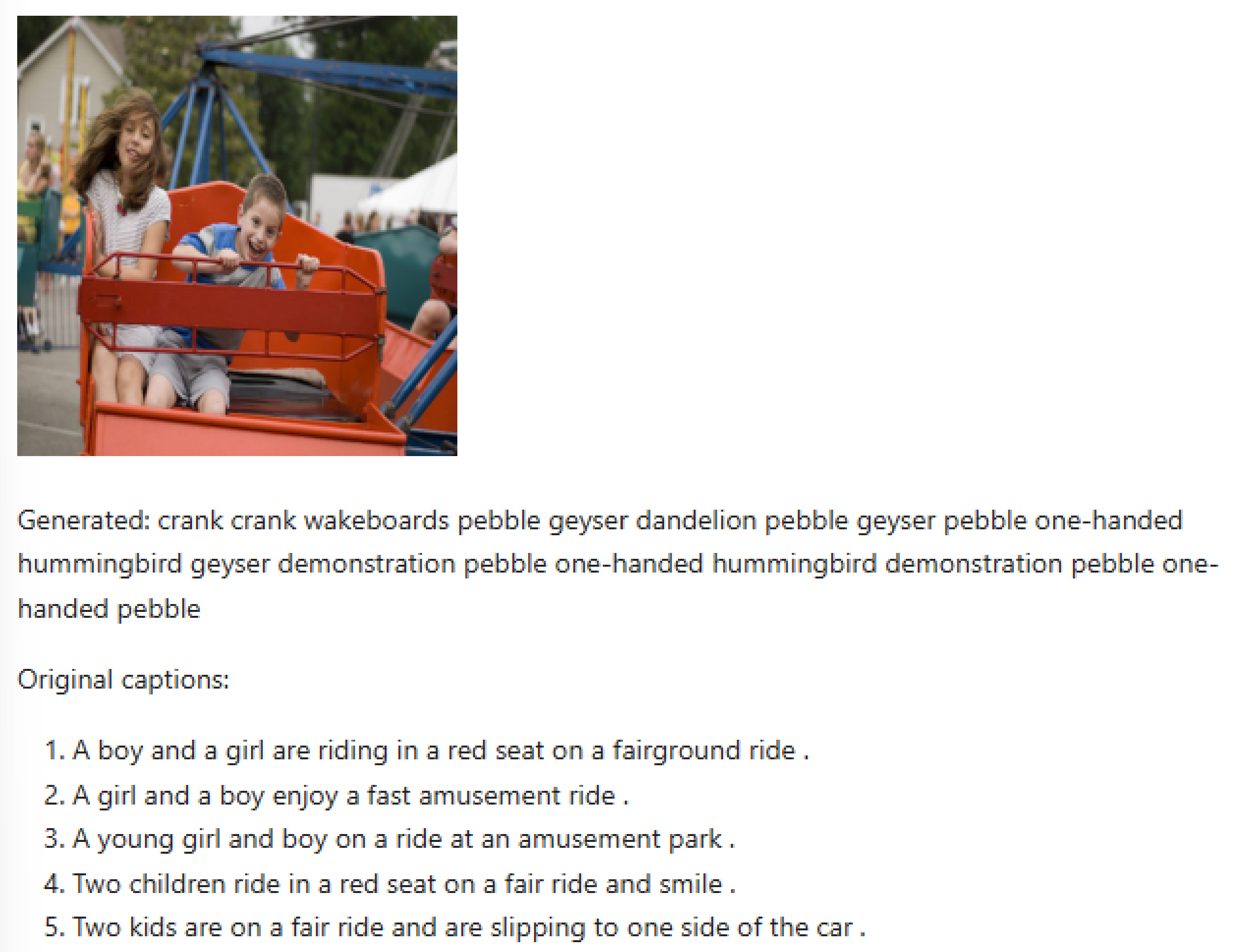
BLEU-1 Scores by Model

BLEU-2 Scores by Model

BLEU-3 Scores by Model

BLEU-4 Scores by Model

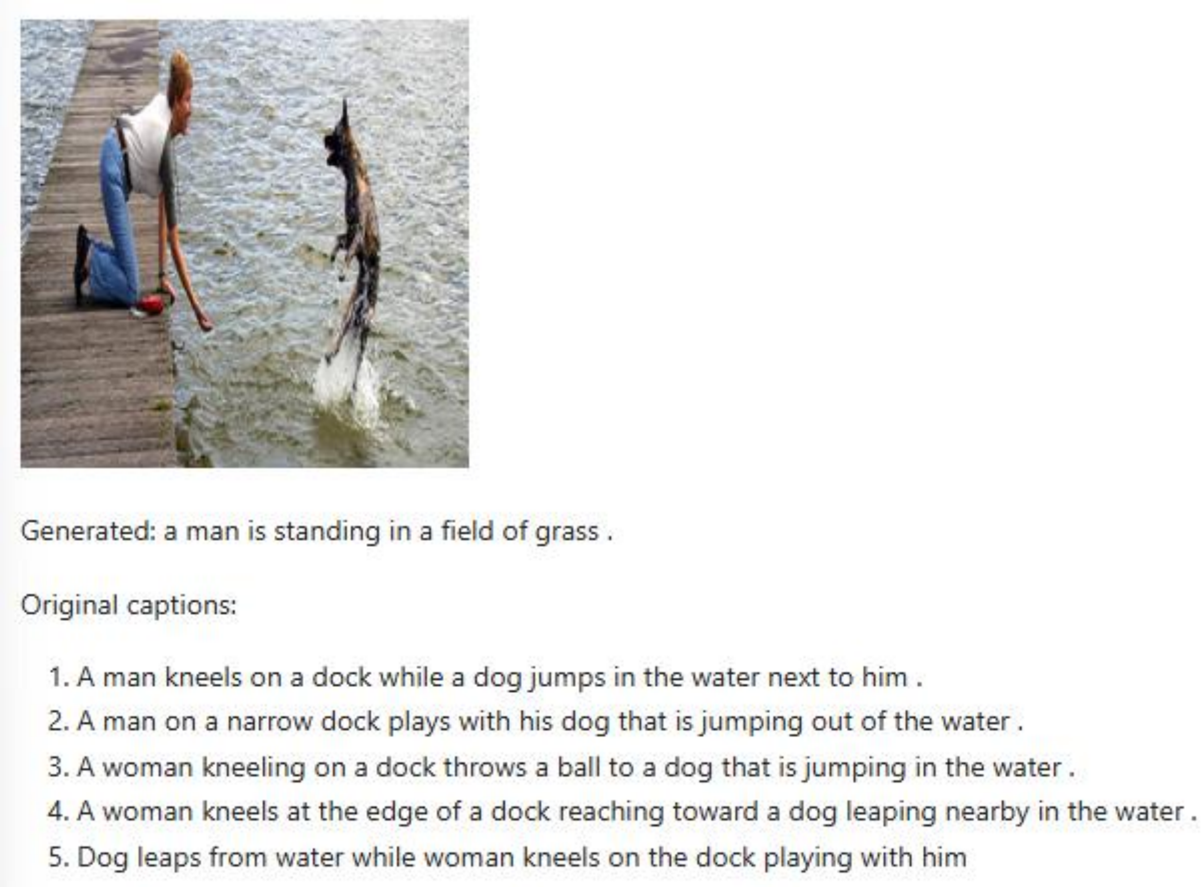
Evaluation Show and Tell - Greedy



Generated: crank crank wakeboards pebble geyser dandelion pebble geyser pebble one-handed hummingbird geyser demonstration pebble one-handed hummingbird demonstration pebble one-handed pebble

Original captions:

1. A boy and a girl are riding in a red seat on a fairground ride .
2. A girl and a boy enjoy a fast amusement ride .
3. A young girl and boy on a ride at an amusement park .
4. Two children ride in a red seat on a fair ride and smile .
5. Two kids are on a fair ride and are slipping to one side of the car .

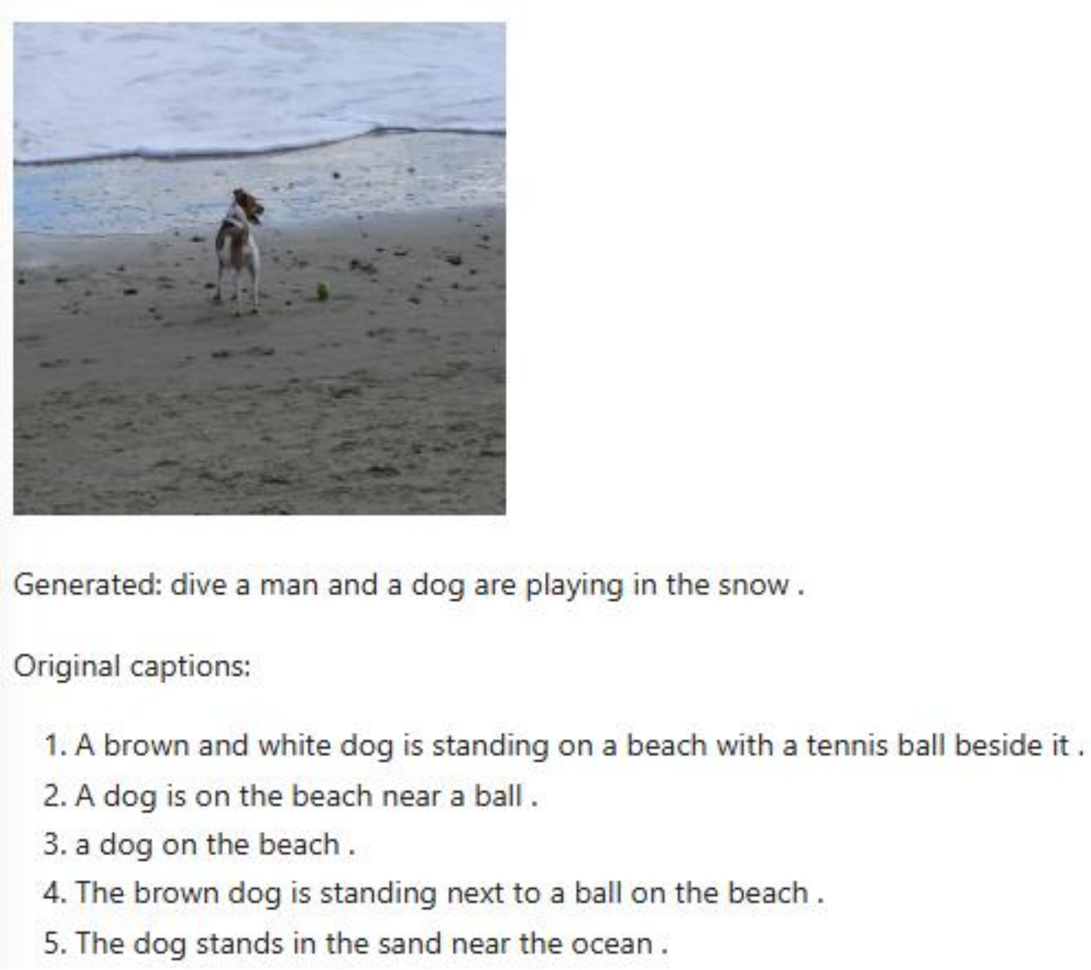


Generated: a man is standing in a field of grass .

Original captions:

1. A man kneels on a dock while a dog jumps in the water next to him .
2. A man on a narrow dock plays with his dog that is jumping out of the water .
3. A woman kneeling on a dock throws a ball to a dog that is jumping in the water .
4. A woman kneels at the edge of a dock reaching toward a dog leaping nearby in the water .
5. Dog leaps from water while woman kneels on the dock playing with him

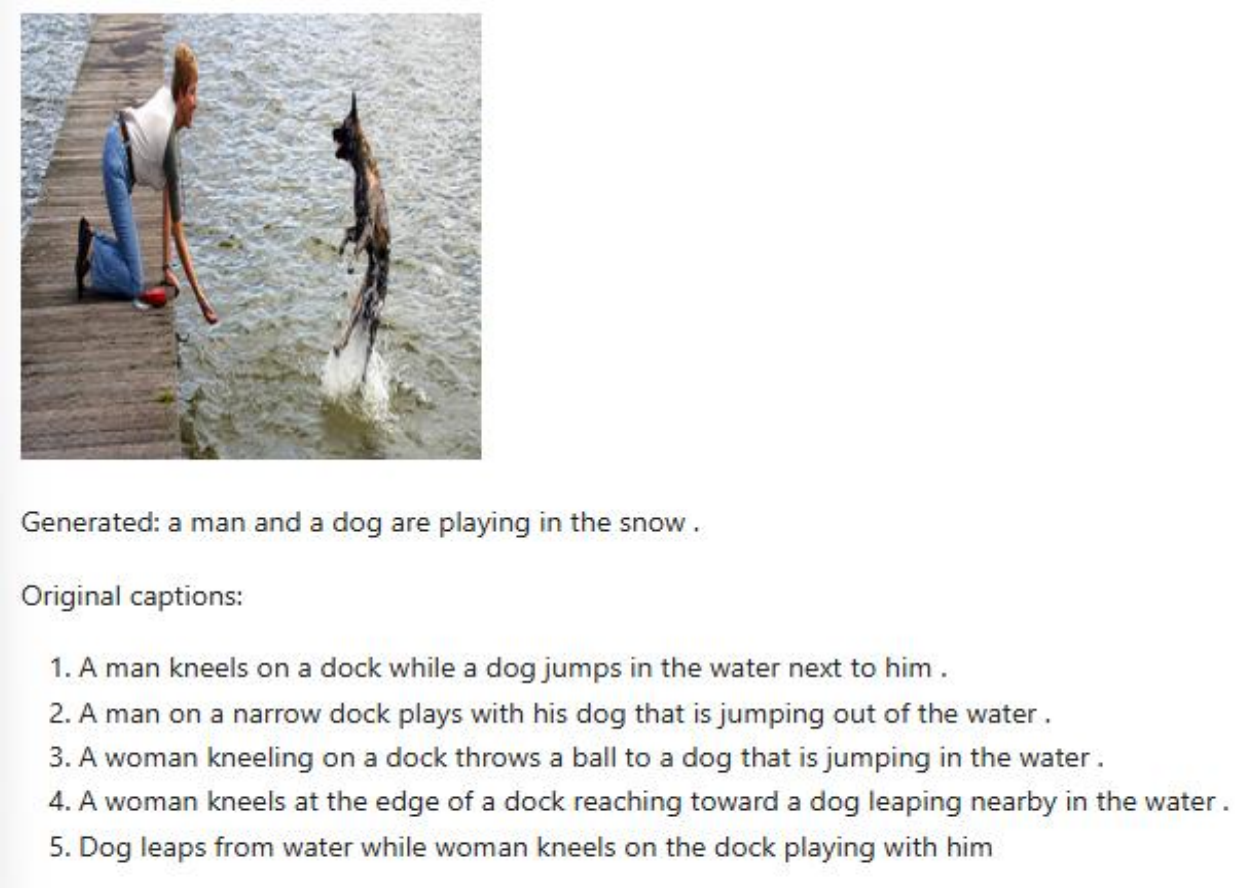
Evaluation Show and Tell - Beam



Generated: dive a man and a dog are playing in the snow .

Original captions:

1. A brown and white dog is standing on a beach with a tennis ball beside it .
2. A dog is on the beach near a ball .
3. a dog on the beach .
4. The brown dog is standing next to a ball on the beach .
5. The dog stands in the sand near the ocean .

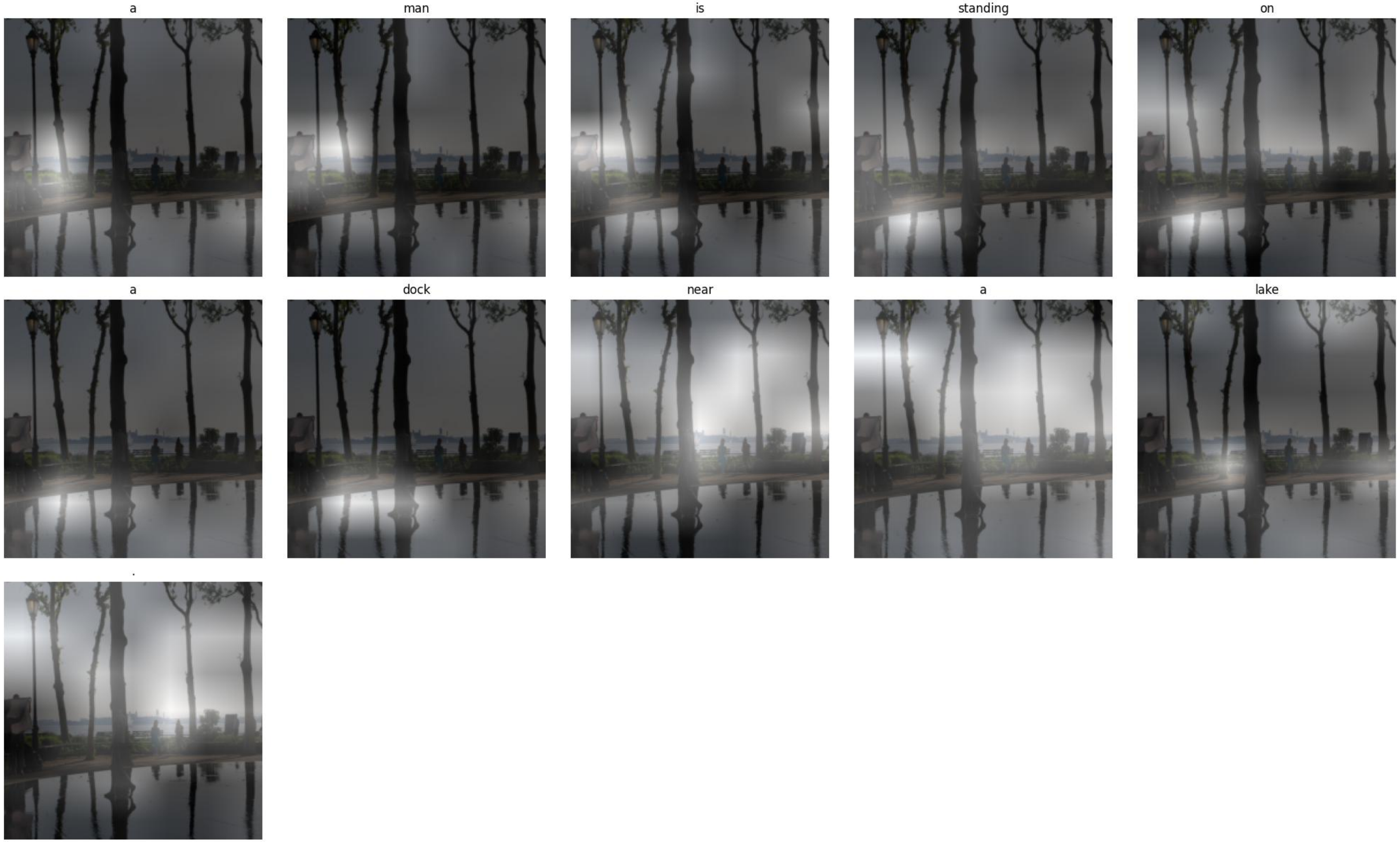


Generated: a man and a dog are playing in the snow .

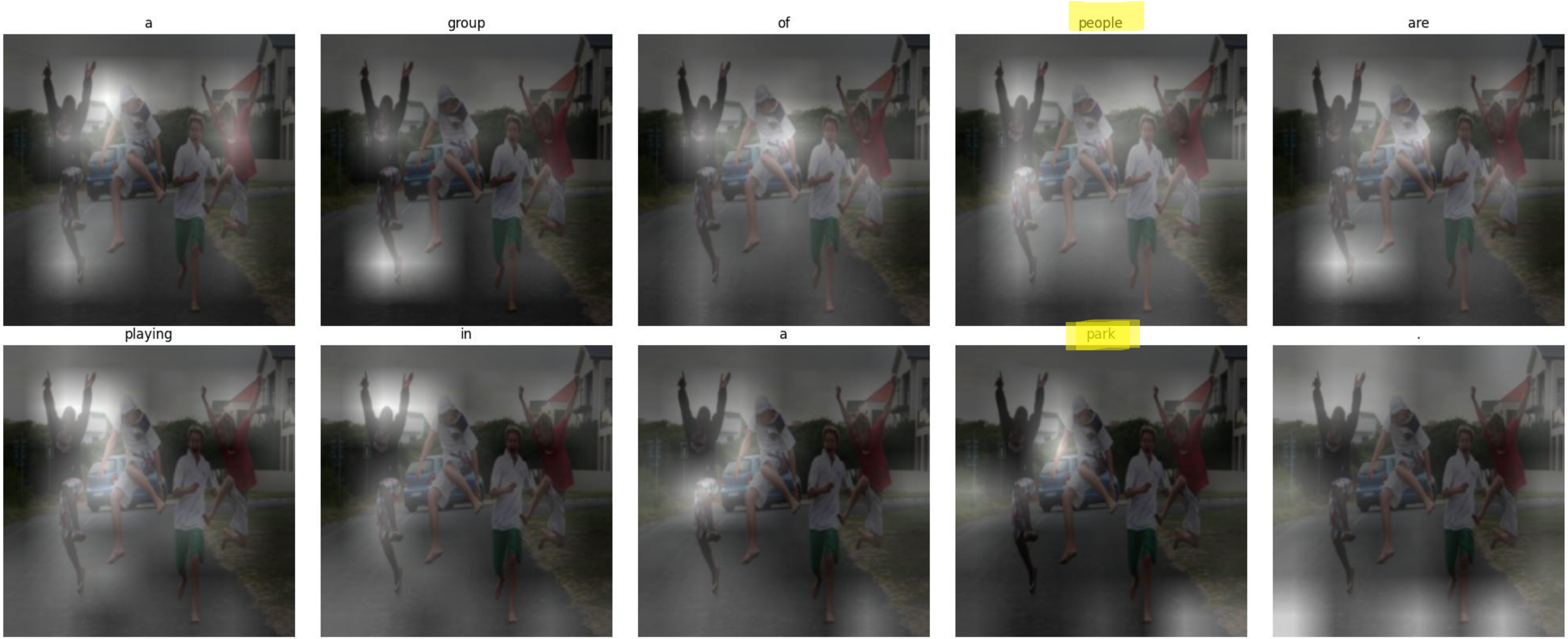
Original captions:

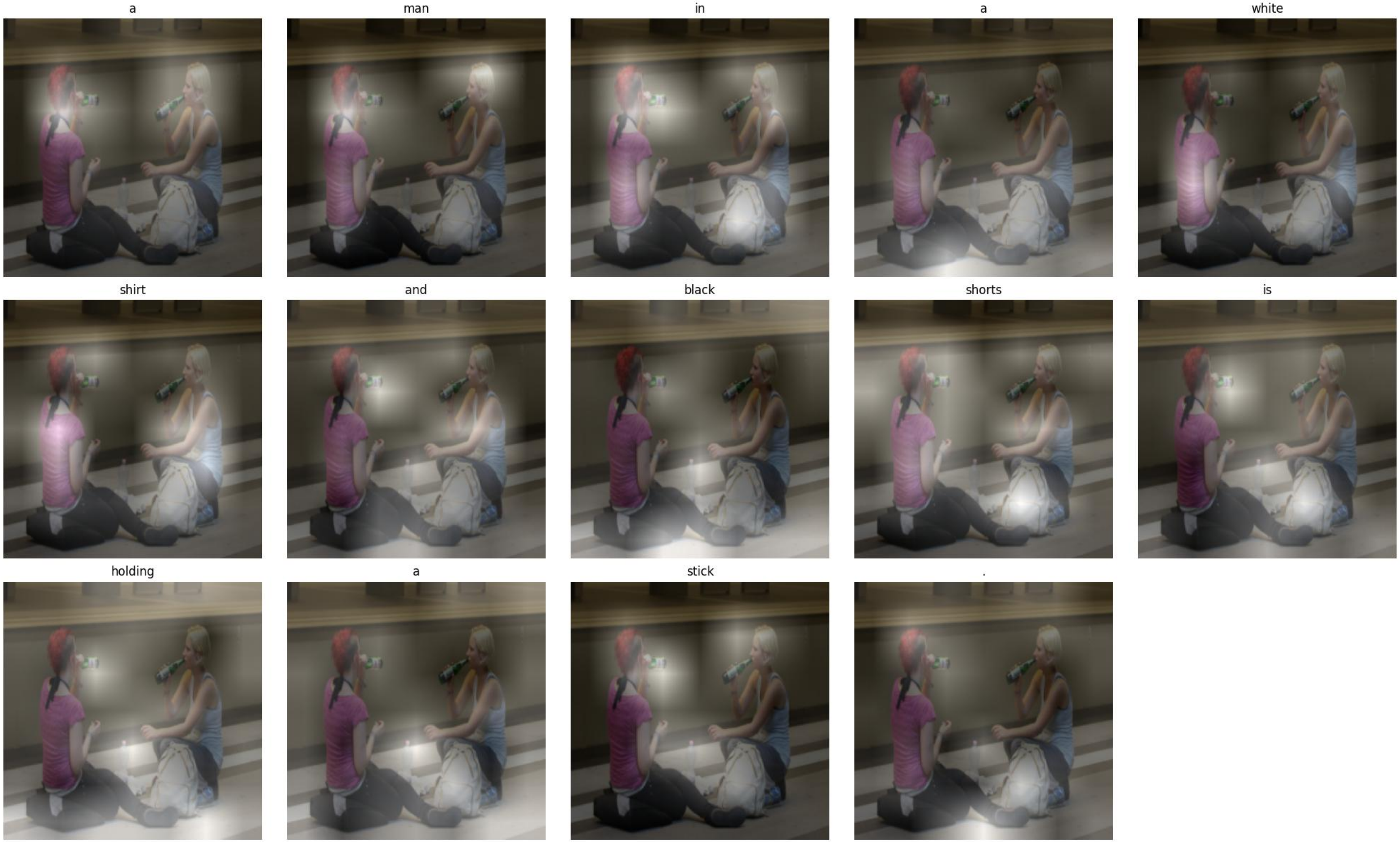
1. A man kneels on a dock while a dog jumps in the water next to him .
2. A man on a narrow dock plays with his dog that is jumping out of the water .
3. A woman kneeling on a dock throws a ball to a dog that is jumping in the water .
4. A woman kneels at the edge of a dock reaching toward a dog leaping nearby in the water .
5. Dog leaps from water while woman kneels on the dock playing with him

Evaluation Show, **Attend** and Tell



Evaluation Show, **Attend** and Tell





Greedy: A man in a blue shirt and a women in a white shirt and black shorts is holding a bat.

Beam: A man in a white shirt and black shorts is holding a stick.

Fazit

- BLEU-Scores von «Show, Attend, and Tell» erreicht
- Attention verhält sich wie erwartet.
- Mögliche Erweiterungen
  - Hard-Attention
  - Fix «Show and Tell»-Modell?
  - Varianten mit anderen pre-trained Encoder und anderen Decodern (z.B. GRU)
  - Hyperparametersuche von Dimensionen (Embedding, Attention, Hidden)