IMAGE CAPTIONING WITH SPARSE LSTM

Yujun Lin
Tsinghua University
linyy14@mails.tsinghua.edu.cn

Song Han
Stanford University
songhan@stanford.edu

Yu Wang
Tsinghua University
yu-wang@tsinghua.edu.cn

William J. Dally
Stanford University
NVIDIA
dally@stanford.edu

ABSTRACT

Long Short-Term Memory (LSTM) is widely used to solve sequence modeling problems, for example, image captioning. We found the LSTM cells are heavily redundant. We adopt network pruning to reduce the redundancy of LSTM and introduce sparsity as new regularization to reduce overfitting. We can achieve better performance than the dense baseline while reducing the total number of parameters in LSTM by more than 80%, from 2.1 million to only 0.4 million. Sparse LSTM can improve the BLUE-4 score by 1.3 points on Flickr8k dataset and CIDER score by 1.7 points on MSCOCO dataset. We explore four types of pruning policies on LSTM and visualize the sparsity pattern, weight distribution of sparse LSTM and analyze the pros and cons of each policy.

1 INTRODUCTION

Automatically describing an image has been an appealing task in recent years. Many innovative, efficient architectures proposed to solve this problem are based on Long Short-Term Memory (LSTM) model for its fascinating effects (Karpathy & Li, 2015; Vinyals et al., 2016; Xu et al., 2015). However, memory bandwidth of the hardware affects inference performance of LSTM, since the matrix-vector multiplication is memory bounded. Along with the storage and energy cost, it brings great challenges to mobile deployment. As claimed in Han et al. (2015) and See et al. (2016), network pruning has a compelling effect on reducing the number of parameters of neural networks.

Inspired by these works, we utilize network pruning to reduce the size of LSTM networks. At the same time, we introduce sparsity as a regularization to help deal with the overfitting during training. We experimented four simple types of pruning policies; and unlike existing work which prunes a pre-trained model in Han et al. (2015), we merge pre-train and pruning stages to shorten training time and show how this technique works with LSTM networks. We managed to improve the performance with proper pruning policy. Compared with NeuralTalk2(Karpathy & Li, 2015) baseline, on Flickr8k dataset (Cyrus et al., 2010), we improve Bleu-4 score from 18.2 to 19.5. On MSCOCO dataset (Lin et al., 2014), we improve CIDER score from 91.4 to 93.1.

2 SPARSE TRAINING ON LSTM NETWORKS

The core of training sparse LSTM networks is pruning: removing small connections. During model initialization, a binary mask full of one is allocated to each weight in LSTM. Training still calls the dense linear algebra library, but we multiply the weight with its corresponding mask after optimizer update step in every iteration. The sparsity, i.e. the proportion of zeros in the mask, determines how many connections are trimmed.

We explore the following choices of the pruning policy: whether sparsity is consistent or not, when to start pruning, when to update masks and based on sort algorithm (Han et al., 2016b) or threshold algorithm (Narang et al., 2017). Simple pruning policies can be divided into four types as shown in Figure 1. Type I is to prune the weights from the second iteration with fixed sparsity. Type II is similar to Type I except that it starts from the second epoch. Type III begins pruning from the second epoch and increases the sparsity step by step while training. All these three types of policies update masks every iteration based on the sort algorithm. Type IV uses a continuously increasing threshold to discard smaller weights and updates masks at regular intervals.

Sort Algorithm	**Threshold Algorithm**
$S = \text{sort}(\lvert\mathbf{W}^{(t)}\rvert)$	if $(si < t < ei$ and $t \,\%\, f == 0)$:
$\lambda = S_k$	$\lambda = s \times (t - si + 1) \,/\, f$
$Mask = \mathbb{1}(\mathbf{W}^{(t)} > \lambda)$	$Mask = \mathbb{1}(\mathbf{W}^{(t)} > \lambda)$
$\mathbf{W}^{(t)} = \mathbf{W}^{(t)} \cdot Mask$	

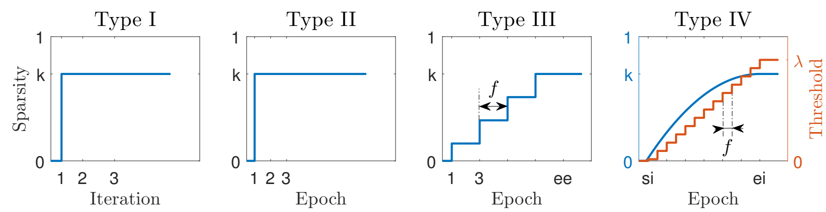


Figure 1: Four types of pruning policies

Type I and Type II have only one hyper-parameter, the sparsity k. There are two more hyper-parameters in Type III, the period f to increase sparsity and the epoch ee to stop growing sparsity. Type IV has four hyper-parameters: si is the iteration to start pruning; ei is the iteration to end updating masks; f is the period to update masks; s is the slope to increase threshold.

3 Experiments

We run all our experiments with NeuralTalk2 (Karpathy & Li, 2015). It uses a CNN (VGG net) as feature extractor, and the following LSTM network to generate image descriptions. For evaluation, we adopt two widely used datasets: Flickr8k (Cyrus et al., 2010) and MSCOCO (Lin et al., 2014). We select separately two sets of 5K random images from MSCOCO validation dataset as our validation and test sets and use them to report results in the following section. The baseline model we used is trained with default hyper-parameters[1]. It achieved BLEU 1-4 score of [59.3, 40.8, 27.2, 18.2] on Flickr8k test dataset and a CIDER score of 91.4 on MSCOCO test dataset. To explore the impact of pruning on LSTM, we only prune the weights in the LSTM, not the CNN.

Figure 2 shows in detail the learning curves on MSCOCO dataset with sparsity 80%. For Type I and II, the damage on performance brought by pruning is recovered by fine-tuning CNN feature extractor. For Type III, the final rise of sparsity results in a surge of validation loss but soon the loss falls near the baseline. As training goes further, validation loss curves of sparse models decline slowly to a slightly lower level than baseline. It shows that the sparsity of weight matrix contributes to reducing overfitting as regularization.

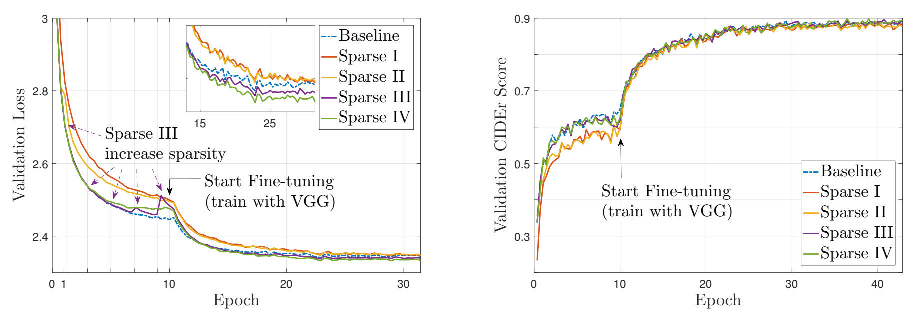


Figure 2: The validation loss and CIDER score curves for the baseline and sparse training on MSCOCO dataset. The sparsity is 80% (20% non-zero) for all sparse cases.

3.1 Image Caption Comparison

Table 1 reports the results on two datasets when we pruned 80% of origin LSTM size. Since BLEU score is not a perfect metric, we also visualize some images and corresponding captions in Flickr8k test dataset in Figure 3. Pruning mainly improves originally-worse captions in the baseline, like image 1, 4, 5. More image caption results generated by four types of sparse LSTMs are provided in the appendix. Therefore, although the capacity of LSTM is significantly reduced, the model itself is not badly damaged and somewhat improves its performance. We can roughly have a performance ranking: Type I > Type II $\approx$ Type IV > Type III $\approx$ Baseline.

[1]The default hyper-parameters are provided on `https://github.com/karpathy/neuraltalk2`.

Table 1: Results on Neuraltalk2 (Sparsity: 80%)

Type	Flickr8k				MSCOCO			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	BLEU-1	BLEU-4	METEOR
Baseline	59.3	40.8	27.2	18.2	91.4	71.1	29.5	24.2
Sparse Type I	60.4 (+ 1.1)	42 (+ 1.2)	28.6 (+ 1.4)	**19.5** (+ 1.3)	**93.1** (+ 1.7)	**72.0** (+ 0.9)	**30.1** (+ 0.6)	**24.3** (+ 0.1)
Sparse Type II	60.4 (+ 1.1)	41.7 (+ 0.9)	28.0 (+ 0.8)	18.8 (+ 0.6)	92.6 (+ 1.2)	71.7 (+ 0.6)	29.8 (+0.3)	24.2 (+ 0)
Sparse Type III	**60.7** (+ 1.4)	**42.3** (+ 1.5)	**28.9** (+ 1.7)	18.6 (+ 0.4)	91.5 (+ 0.1)	71.5 (+ 0.4)	29.5 (+ 0)	24.1 (- 0.1)
Sparse Type IV	60.2 (+ 0.9)	42.0 (+ 1.2)	28.5 (+ 1.3)	18.9 (+ 0.7)	92.3 (+ 0.9)	71.6 (+ 0.5)	29.6 (+ 0.1)	24.2 (+ 0)

✗Baseline: a group of children are playing in a fountain
+Type I: three children are playing in a park
+Type II: three children are playing in a park
✓Type III: a group of children playing in a park
✓Type IV: a group of children are playing in a park

✓Baseline: a white and brown dog is jumping up to catch a frisbee
−Type I: a white and brown dog is jumping over a hurdle
−Type II: a white dog is jumping over a hurdle
−Type III: a white and brown dog is jumping over a hurdle
−Type IV: a white and brown dog is jumping over a hurdle

✓Baseline: a dog jumps over a hurdle
✓Type I: a dog jumps over a hurdle
✓Type II: a dog jumps over a hurdle
✓Type III: a dog jumps over a hurdle
✓Type IV: a brown and white dog jumps over a hurdle

✗Baseline: a group of people are in the air in front of a crowd
✓Type I: a group of people are playing basketball
✓Type II: a group of people are playing basketball
+Type III: a man in a blue shirt is jumping off a basketball court
✓Type IV: a group of people are playing basketball in a stadium

−Baseline: two young boys are playing with a soccer ball
✓Type I: two boys play in a park
✓Type II: two boys play in the grass
✓Type III: two children are playing in a field
+Type IV: two children are playing in a yard

Figure 3: Visualization of image captions generated by four sparse LSTM models.

3.2 Visualizing the Sparsity Pattern

We visualize the sparsity pattern of LSTM and can easily distinguish the four gates of LSTM (i, f, o, g) in Figure 4. Different gates are allocated with different sparsity ratio. We also observe that the dynamic range of weight values will expand after pruning. This is reasonable because pruning reduces the number of connections between layers and thus the neural network enhances valuable connections while training. In sorting-based pruning, the shape of the main lobe of weight distribution remains similar to baseline except for a deep notch around zero value. In Type IV which is based on threshold algorithm, pronounced bimodality appears with less expansion on the range of weights than others. These sharp and concentrated peaks help further quantization for network compression (Han et al., 2016a).

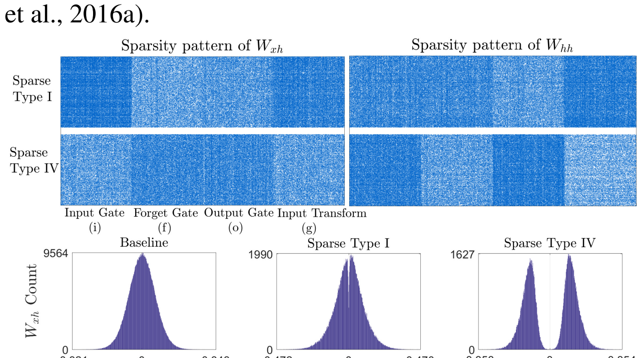


Figure 4: Visualization of sparsity pattern and weight distribution in LSTM. (Sparsity: 80%)

4 Conclusion

We explored four types of pruning policy for image caption LSTM to decrease the model size and improve quality. Type I and II have only one hyper-parameter and thus are the easiest way to reduce the number parameters in LSTM and can even achieve better performance on image captioning. Type I can improve BLEU-4 score by 1.3 points on Flickr8k dataset and CIDER score by 1.7 points on MSCOCO dataset. Both Type III and IV can maintain baseline performance, but they require more dedicated work on designing hyper-parameters. We also show that pruning weights in LSTM during the initial training will save training time compared with fine-tuning on the pre-trained model.

REFERENCES

Rashtchian Cyrus, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *NAACL HLT Workshop*, pp. 139–147. Association for Computational Linguistics, 2010.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *NIPS*, 2015.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016a.

Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Shijian Tang, Erich Elsen, Bryan Catanzaro, John Tran, and William J. Dally. DSD: regularizing deep neural networks with dense-sparse-dense training flow. *CoRR*, abs/1607.04381, 2016b. URL `http://arxiv.org/abs/1607.04381`.

Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Sharan Narang, Greg Diamos, Shubho Sengupta, and Erich Elsen. Exploring sparsity in recurrent neural networks. In *ICLR*, 2017.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. Compression of neural machine translation models via pruning. *CoRR*, abs/1606.09274, 2016. URL `http://arxiv.org/abs/1606.09274`.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

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. *ICML*, 2015.

A Appendix: More Examples of Image Captioning after Four Types Pruning on NeuralTalk2 (Images from Flickr8k Test Set)

✗ Unrelated to the image — Somewhat related to the image + Describes with minor error ✓ Describes without errors

✓Baseline: a young boy jumping into a pool
✓Sparse Type I: a young boy is jumping into a swimming pool
+Sparse Type II: a young boy is playing in a pool
✓Sparse Type III: a young boy jumping in to a pool
+Sparse Type IV: a boy in a blue shirt is jumping into a pool

+Baseline: a group of people are in a raft on a lake
✓Sparse Type I: a group of people are sitting on a raft in the water
✓Sparse Type II: a group of people are riding on a raft
✓Sparse Type III: a group of people are riding on a raft
✓Sparse Type IV: a group of people are sitting on a raft in the water

✓Baseline: a man climbing a rock wall
✓Sparse Type I: a man climbing a rock wall
✓Sparse Type II: a man climbing a rock wall
✓Sparse Type III: a man is climbing a rock face
✓Sparse Type IV: a man in a red shirt climbs a rock wall

—Baseline: a dog is running through the water
✓Sparse Type I: a dog is jumping into the water
✓Sparse Type II: a brown dog is jumping into the water
✓Sparse Type III: a brown dog is jumping into the water
✗Sparse Type IV: a man is standing on top of a cliff overlooking a lake

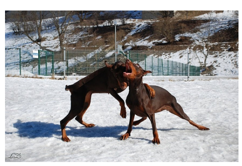
+Baseline: two black dogs run through a field
✓Sparse Type I: a black dog is running through a grassy field
✓Sparse Type II: a black dog is running through a grassy field
✓Sparse Type III: a black dog is running through the grass
✓Sparse Type IV: a black dog is running through a grassy field

+Baseline: a boy in a red shirt is swinging on a swing
✓Sparse Type I: a little boy in a blue shirt is swinging on a swing
+Sparse Type II: a little boy in a blue shirt is swinging on a swing
+Sparse Type III: a boy in a red shirt is swinging on a swing
✓Sparse Type IV: a child in a blue shirt is swinging on a swing

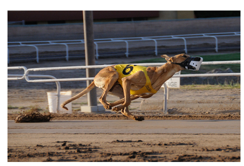
—Baseline: a boy in a blue shirt is jumping off of a swing
✓Sparse Type I: a boy in a blue shirt is jumping on a trampoline
✓Sparse Type II: a boy in a blue shirt is jumping on a trampoline
✗Sparse Type III: a boy in a red shirt is jumping off a wooden ramp
✓Sparse Type IV: a boy in a blue shirts jumping on a trampoline

—Baseline: a white dog is running on the beach
✓Sparse Type I: a white dog is running through the water
✓Sparse Type II: a white dog is running through the water
✓Sparse Type III: a white dog is running through the water
✓Sparse Type IV: a white dog runs through the water

—Baseline: a young boy is jumping into a swimming pool
✓Sparse Type I: a boy in a swimming pool
—Sparse Type II: a young boy is jumping into a swimming pool
✓Sparse Type III: a little boy is playing in a pool
✓Sparse Type IV: a little boy in a swimming pool

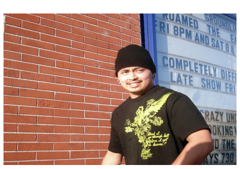


—Baseline: two dogs are running through the snow
✓Sparse Type I: a brown dog and a black dog are playing in the snow
—Sparse Type II: a brown dog is running through the snow
—Sparse Type III: two dogs are running in the snow
✓Sparse Type IV: two dogs are playing in the snow



✗Baseline: a brown dog is jumping over a hurdle
✗Sparse Type I: a brown dog is jumping over a hurdle
+Sparse Type II: a brown and white dog is running on a track
✓Sparse Type III: a brown dog is running on a track
✓Sparse Type IV: a brown dog is running on a track

—Baseline: a group of people are standing in front of a building
✓Sparse Type I: a group of people are posing for a picture
✓Sparse Type II: a group of people are posing for a picture
✓Sparse Type III: a group of people are posing for a picture
—Sparse Type IV: a group of people are standing in front of a building



✗Baseline: a man in a black jacket is standing next to a man in a red shirt
✗Sparse Type I: a man in a black jacket is standing in front of a white building
✗Sparse Type II: a man in a black shirt and a white shirt is standing in front of a
+Sparse Type III: a man in a black jacket is standing in front of a brick wall
+Sparse Type IV: a man in a black jacket is standing in front of a brick wall

—Baseline: a young boy in a red shirt is playing with a soccer ball
✗Sparse Type I: a young boy in a red shirt and blue jeans is walking on a sidewalk
+Sparse Type II: a boy in a red shirt is standing on a sidewalk
✓Sparse Type III: two children are playing in the street
✓Sparse Type IV: two young children are playing in a park

—Baseline: a young boy wearing a red shirt and blue jeans is running on a street
—Sparse Type I: a young boy wearing a blue shirt and blue jeans is running on a track
—Sparse Type II: a young boy wearing a red shirt is riding a unicycle
✓Sparse Type III: a young boy wearing a blue shirt and blue jeans is riding a bike
+Sparse Type IV: a boy in a red shirt is riding a bike

✗Baseline: a man in a blue wetsuit is surfing
✓Sparse Type I: a young boy in a red shirt is playing in the water
+Sparse Type II: a young boy wearing a blue shirt and blue jeans is standing in the water
✓Sparse Type III: a young boy wearing a red shirt is standing in the water
✓Sparse Type IV: a young boy in a red shirt is playing in the water

✗ Unrelated to the image — Somewhat related to the image + Describes with minor error ✓ Describes without errors

✗Baseline: a woman in a white shirt and black pants is standing in front of a crowd
✗Sparse Type I: a woman wearing a white shirt and a black hat is standing in front of a
✗Sparse Type II: a woman in a white shirt and black pants is standing in front of a
+Sparse Type III: a woman in a blue shirt and sunglasses smiles
—Sparse Type IV: a woman wearing a black shirt and white pants is standing on the sidewalk

—Baseline: a young boy wearing a red shirt is smiling.
✓Sparse Type I: a young boy with a blue shirt and a blue shirt
+Sparse Type II: a young boy wearing a blue shirt is sitting on a UNK
✓Sparse Type III: a young boy wearing a blue shirt is looking at the camera
✓Sparse Type IV: a young boy is wearing a blue shirt

✗Baseline: a group of people are standing on a snowy hill
+Sparse Type I: a person in a yellow car is driving through the water
—Sparse Type II: a group of people are walking through a river
✓Sparse Type III: a person in a blue car is driving through the water
+Sparse Type IV: a person in a yellow car is driving through the water

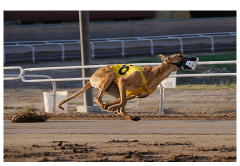
✗Baseline: a brown dog is digging in the sand
+Sparse Type I: a brown dog with a red collar is standing in the water
+Sparse Type II: a brown dog with a collar is standing in the water
+Sparse Type III: a brown dog is running through the water
+Sparse Type IV: a brown dog is standing in the water

—Baseline: a man in a black jacket is standing next to a man in a red shirt
—Sparse Type I: a man in a black jacket is standing in front of a white building
+Sparse Type II: a man in a black jacket is standing in front of a brick wall
+Sparse Type III: a man in a black jacket is standing in front of a brick wall
+Sparse Type IV: a man in a black jacket is standing in front of a brick wall

—Baseline: a group of people are standing in front of a white building
+Sparse Type I: a group of children are playing in a park
✓Sparse Type II: a group of children are playing in a field
—Sparse Type III: a group of children playing soccer
✓Sparse Type IV: a group of people are standing in a field

—Baseline: a girl in a pink shirt is jumping off a rock into the air
+Sparse Type I: a girl in a pink shirt is jumping on a trampoline
+Sparse Type II: a girl in a pink shirt is jumping on a trampoline
—Sparse Type III: a girl in a pink shirt is jumping into a pool
+Sparse Type IV: a girl in a pink shirt is jumping on a trampoline

+Baseline: two dogs are running through the snow
+Sparse Type I: two dogs running in the snow
✓Sparse Type II: a black and white dog is running through the snow
✓Sparse Type III: a black and white dog is running through the snow
✓Sparse Type IV: a black and white dog is running through the snow

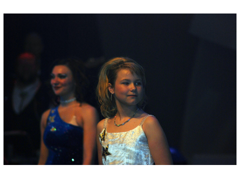


✗Baseline: a brown dog is jumping over a hurdle
✗Sparse Type I: a brown dog is jumping over a hurdle
+Sparse Type II: a brown and white dog is running on a track
✓Sparse Type III: a brown dog is running on a track
✓Sparse Type IV: a brown dog is running on a track

✗Baseline: a man and a woman sit on a bench outside a building
✓Sparse Type I: a man in a black shirt is standing next to a man in a black shirt
✓Sparse Type II: a man in a black shirt is standing in front of a large building
✗Sparse Type III: a man and a woman are sitting on a bench
✓Sparse Type IV: a man in a black shirt and jeans is standing on a sidewalk

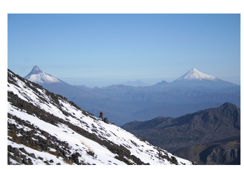
✗Baseline: a woman in a bikini is standing in front of a waterfall
✗Sparse Type I: a woman in a bikini is standing in front of a waterfall
+Sparse Type II: a woman in a black and white shirt is standing in front of the ocean
✓Sparse Type III: two young girls are standing on a beach
+Sparse Type IV: a woman in a blue shirt is standing on a beach

✓Baseline: a group of men play basketball
✓Sparse Type I: a group of men play basketball
✓Sparse Type II: a basketball player in a white uniform is dribbling the ball
—Sparse Type III: a basketball player in a white uniform is trying to score
✓Sparse Type IV: a group of men play basketball



—Baseline: a woman in a black shirt and a woman in a black dress
+Sparse Type I: a woman in a white shirt and a girl is smiling
—Sparse Type II: a woman in a white dress is holding a microphone
+Sparse Type III: a woman in a blue shirt and a woman in a white dress
✗Sparse Type IV: a woman in a black shirt and a white hat is smiling

—Baseline: a brown and white dog is playing with a red ball
✓Sparse Type I: a brown and white dog is running on a grassy field
✓Sparse Type II: a brown and white dog is running on the grass
✓Sparse Type III: a brown and white dog is running through a grassy field
✓Sparse Type IV: a brown and white dog is running on a grassy field



+Baseline: a man is standing on top of a snowy mountain
✓Sparse Type I: a man is standing on a snowy mountain
+Sparse Type II: a man is standing on top of a snowy mountain
✓Sparse Type III: a man is standing on a snowy mountain
✓Sparse Type IV: a man stands on a snowy mountain

✗Baseline: a group of people are standing in front of a crowd
+Sparse Type I: a group of people are standing in front of a building
+Sparse Type II: a group of people are standing in front of a building
+Sparse Type III: a man in a red shirt is standing in front of a building
+Sparse Type IV: a group of people are standing in front of a building