OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks

Presentation by Amun Kharel
CS 6804: Multimodal Vision
Spring 2023
Instructor: Dr. Chris Thomas
Attention Mechanisms

timeseries data \( x = [x_0 \ldots x_n] \)

re-weighing \( \mathbf{w}_i = [w_{i0} \ldots w_{in}] \)

\( y_i = \mathbf{w}_i \cdot x \)

\( y_{i+1} = \mathbf{w}_{i+1} \cdot x \)

is more useful because datapoints gave each other context.

https://www.youtube.com/watch?v=yGTUuEx3GkA&ab_channel=Rasa
Self-Attention

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. CoRR abs/1706.03762, (2017). Retrieved from http://arxiv.org/abs/1706.03762
Self-Attention

https://www.youtube.com/watch?v=yGTUuEx3GkA&ab_channel=Rasa
Application of BERT

- Text Encoding
- Response Selection
- Text Summarization
- Question Answering
- Similarity Retrieval
- And More...

https://jalammar.github.io/illustrated-bert/
High-Level Overview of BERT

https://jalammar.github.io/illustrated-bert/
High-Level Overview of BERT

https://jalammar.github.io/illustrated-bert/
Simple Search Engine Using BERT

https://jalammar.github.io/illustrated-bert/
Simple Search Engine Using BERT

https://jalammar.github.io/illustrated-bert/
Simple Search Engine Using BERT

https://jalammar.github.io/illustrated-bert/
Use Cases of BERT

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

- Model: (pre-trained in step #1)
- Dataset:
- Objective: Predict the masked word (language modeling)

| Email message | Class |
|---------------|-------|
| Buy these pills | Spam  |
| Win cash prizes | Spam  |
| Dear Mr. Almires, please find attached... | Not Spam |

https://jalammar.github.io/illustrated-bert/
BERT architecture

https://jalammar.github.io/illustrated-bert/
BERT architecture

https://jalammar.github.io/illustrated-bert/
BERT Classifier

https://jalammar.github.io/illustrated-bert/
BERT: Masked Language Model

Use the output of the masked word’s position to predict the masked word.

Possible classes:
- All English words
- Improvisation
- Zyzyva

Randomly mask 15% of tokens

Input

https://jalammar.github.io/illustrated-bert/
BERT: Next Sentence Prediction

Predict likelihood that sentence B belongs after sentence A

1% IsNext
99% NotNext

FFNN + Softmax

Tokenized Input

Input

https://jalammar.github.io/illustrated-bert/
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Which NFL team represented the AFC at Super Bowl 50?

Ground Truth Answers: Denver Broncos, Denver Broncos, Denver Broncos

Prediction: Denver Broncos
BERT: Question Answering

Input Preparation

**Question:** How many parameters does BERT-large have?

**Reference Text:** BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

https://www.youtube.com/watch?v=l8ZYCvgGu0o&ab_channel=ChrisMcCormickAI
BERT: Question Answering

This length 768 vector is the weights for the start token classifier. The same weights are applied to every position.

https://www.youtube.com/watch?v=l8ZYCvgGu0o&ab_channel=ChrisMcCormickAI
BERT: Question Answering

https://www.youtube.com/watch?v=l8ZYCvgGu0o&ab_channel=ChrisMcCormickAI
Previous work simply concatenate image region features and text features to learn image-text semantic alignments in a brute force manner.
OSCAR: Problem Addressed by Paper (Pre-training)

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/
OSCAR: Problem Addressed by Paper

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/
OSCAR: Solution Offered

- Word-Tag-Region Triplet

Word-Tag-Region Triplet

\[
\text{A dog is sitting on a couch}, \quad \text{Dog, Couch}
\]

Retrieved from https://arxiv.org/abs/2004.06165, (2020).
OSCAR: Solution Offered (Pre-training)

- Training on Modality View using Contrastive Loss
- Training on Dictionary View Using Masked Token Loss

![Diagram of OSCAR model](Diagram.png)

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from [https://arxiv.org/abs/2004.06165](https://arxiv.org/abs/2004.06165)
OSCAR: Solution Offered (Fine-tuning)

### Understanding
- VQA
- GQA
- NLVR2
- Image-Text Retrieval
- Text-Image Retrieval

### Generation
- Image Captioning
- Novel Object Captioning

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
Motivation: Pre-training

Vision
- Object Detection
- Instance Segmentation
- Semantic Segmentation

Language
- BERT (Devlin et al, 2018)
  - Question Answering
  - Natural Language Inference
  - Sentence Classification

Vision & Language
- Google Conceptual Captions Dataset (CC)
  - Visual Question Answering
  - Image/Text Retrieval
  - Image Captioning

http://valser.org/webinar/slide/slides/20210908%E7%9F%AD%E6%95%99%E7%A8%8B03--%E8%A7%86%E8%A7%89%E4%B8%8E%E8%AF%AD%E8%A8%80%E6%99%BA%E8%83%BD/Lecture2_transformer_and_vlpretraining.pdf
## Motivation: Vision Language Tasks

| Input | Text-to-Image Retrieval | Image-to-Text Retrieval | VQA | Image Captioning | Text-to-Image Generation |
|-------|-------------------------|-------------------------|-----|-------------------|--------------------------|
|       | Query: A couple of zebra walking across a dirt road. | Query: A pool of images. | Image: A couple of zebra walking across a dirt road. | Q: why did the zebra cross the road? | Image: A couple of zebra walking across a dirt road. | Text: A couple of zebra walking across a dirt road. |
| Output | A couple of zebra walking across a dirt road. | A pool of texts. | A: to get to the other side (Selected from a pool of 3,129 answers in VQAv2) | Generation | A couple of zebra walking across a dirt road. | Generation |

http://valser.org/webinar/slide/slides/20210908%E7%9F%AD%E6%95%99%E7%A8%8B03--%E8%A7%86%E8%A7%89%E4%B8%8E%E8%AF%AD%E8%A8%80%E6%99%BA%E8%83%BD/Lecture1_representations_and_attentions.pdf
Related Works: Image Captioning Evolution (Traditional)

1) Object(s)/Stuff
   a) dog
   b) person
   c) sofa

2) Attributes
   brown 0.01
   striped 0.16
   furry 0.26
   wooden 2
   feathered 0.06
   ...

3) Prepositions
   near(s, t) 1
   near(t, u) 1
   against(u, v) 3
   against(v, t) 0.04
   beside(s, t) 1.24
   beside(t, u) 1.17
   ...

4) Constructed CRF

5) Predicted Labeling
   <<null, person_b>, against, <<brown, sofa_c>>
   <<null, dog_a>, near, <<null, person_b>>
   <<null, dog_a>, beside, <<brown, sofa_c>>

6) Generated Sentences
   This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

Baby Talk: Understanding and Generating Image Descriptions. Kulkarni et al., CVPR, 2011

https://yuxng.github.io/Courses/CS6384Spring2022/lecture_25_images_languages.pdf
Related Works: Image Captioning Evolution (RNNs)

(A. Karpathy and L. Fei-Fei, 2015)

(Lu et al., 2018)

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/
Related Works: Image Captioning Evolution (Attention)

Image Captioning with Attentions

1. Input Image  
2. Convolutional Feature Extraction  
3. RNN with attention over the image  
4. Word by word generation

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Xu et al., PMLR, 2015.

https://yuxng.github.io/Courses/CS6384Spring2022/lecture_25_images_languages.pdf
Related Works: Image Captioning Evolution (Current)

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/
Problem with the Current Work

❖ Ambiguity
  ➢ Visual Region features are extracted from over-sampled regions via object detectors
  ➢ Overlaps among image regions at different positions

❖ Lack of grounding
  ➢ No label alignments between regions or objects in an image and words or phrase in text
  ➢ Solution: Salient objects in both image and its paired text (anchor points)

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
OSCAR's Approach

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
OSCAR's Approach

(a) Image-text pair  (b) Objects as anchor points  (c) Semantics spaces

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
OSCAR's Approach (Generation of v and q)

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-visual-language-tasks/
OSCAR's Approach (Pre-training Objective)

\[ x \triangleq [ w, q, v ] = [ w, q, v ] \triangleq x' \]

Modality View (Contrastive Loss)

Dictionary View (Masked Token Loss)

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
OSCAR's Approach (Dictionary View)

\[ \mathcal{L}_{MTL} = -E_{(v,h) \sim D} \log p(h_i | h_{\setminus i}, v) \]

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-visual-language-tasks/
OSCAR's Approach (Modality View)

\[ \mathcal{L}_C = -\mathbb{E}_{(h', w) \sim D} \log p(y | f(h', w)) \]

a contrastive loss for the modality view, which measures the model's capability of distinguishing an original triple and its "polluted" version (that is, where an original object tag is replaced with a randomly sampled one).

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
OSCAR's Approach (Full Pre-training Objective)

\[ \mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{C}}. \]
OSCAR's Approach (Implementation Details)

- Two model variants as OSCAR Base ($H = 768$) and OSCAR Large ($H = 1024$)
- Adam Optimizer
- OSCAR Base trained for at least 1.0 M steps with learning rate $5e^{-5}$ and batch size 768
- OSCAR Large trained for at least 900k steps with learning rate $1e^{-5}$ and batch size 512
- Sequence length of discrete token $h$ and region features $v$ are 35 and 50 respectively
OSCAR’s Fine-tuning (Image Captioning)

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-visio
on-language-tasks/
OSCAR's Fine-tuning (Image Captioning Inference)

https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-visual-language-tasks/
OSCAR's Fine-tuning (Image Text Retrieval)

- There are two tasks Image Retrieval and Text Retrieval
- Binary Classification problem using CLS
- Randomly pick different image-text pair and predict if they are aligned or not
- During Test, probability score is used to rank the given image-text pairs of a query
OSCAR's Fine-tuning (Visual Question Answering)

- Model needs to answer using Natural Language questions based on image
- Image and question is given to select answer from multi-choice list
- Concatenate question, object tags and region features
- CLS output is fed for linear classifier for multi-label classification
- Fine-tune model based on cross-entropy loss
- Simply use Softmax function for prediction

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
## Experimental Results and Analysis

| Task  | Image Retrieval R@1 | Image Retrieval R@5 | Image Retrieval R@10 | Text Retrieval R@1 | Text Retrieval R@5 | Text Retrieval R@10 | Image Captioning B@4 | Image Captioning M | Image Captioning C | Image Captioning S | NoCaps C | NoCaps S | VQA test-std | NLVR2 test-P |
|-------|---------------------|---------------------|----------------------|-------------------|-------------------|---------------------|---------------------|-------------------|------------------|------------------|-----------|-----------|--------------|--------------|
| SoTA <sub>S</sub> | 39.2 | 68.0 | 81.3 | 56.6 | 84.5 | 92.0 | 38.9 | 29.2 | 129.8 | 22.4 | 61.5 | 9.2 | 70.90 | 53.50 |
| SoTA <sub>B</sub> | 48.4 | 76.7 | 85.9 | 63.3 | 87.0 | 93.1 | 39.5 | 29.3 | 129.3 | 23.2 | 73.1 | 11.2 | 72.54 | 78.87 |
| SoTA <sub>L</sub> | 51.7 | 78.4 | 86.9 | 66.6 | 89.4 | 94.3 | – | – | – | – | – | – | 73.40 | 79.50 |
| OSCAR <sub>B</sub> | 54.0 | 80.8 | 88.5 | 70.0 | 91.1 | 95.5 | 40.5 | 29.7 | 137.6 | 22.8 | 78.8 | 11.7 | 73.44 | 78.36 |
| OSCAR <sub>L</sub> | 57.5 | 82.8 | 89.8 | 73.5 | 92.2 | 96.0 | 41.7 | 30.6 | 140.0 | 24.5 | 80.9 | 11.3 | 73.82 | 80.37 |
| Δ | 5.8 | 4.4 | 2.9 | 6.9 | 2.8 | 1.7 | 2.2 | 1.3 | 10.7 | 1.3 | 7.8 | 0.5 | 0.42 | 0.87 |

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from [https://arxiv.org/abs/2004.06165](https://arxiv.org/abs/2004.06165)
## Experimental Results and Analysis

| Method       | Size | 1K Test Set | 5K Test Set |
|--------------|------|-------------|-------------|
|              |      | Text Retrieval | Image Retrieval | Text Retrieval | Image Retrieval |
|              |      | R@1  R@5 R@10 | R@1  R@5 R@10 | R@1  R@5 R@10 | R@1  R@5 R@10 |
| DVSA [14]    | -    | 38.4 69.9 80.5 | 27.4 60.2 74.8 | -    | -    | -    | -    |
| VSE++ [7]    | -    | 64.7 95.9 52.0 | 92.0 41.3 81.2 | 30.3 72.4 |
| DPC [46]     | -    | 65.6 98.9 95.5 | 47.1 79.9 90.0 | 41.2 70.5 81.1 | 25.3 53.4 66.4 |
| CAMP [42]    | -    | 72.3 94.8 98.3 | 58.5 87.9 95.0 | 50.1 82.1 89.7 | 39.0 68.9 80.2 |
| SCAN [18]    | -    | 72.7 94.8 98.4 | 58.8 88.4 94.8 | 50.4 82.2 90.0 | 38.6 69.3 80.4 |
| SCG [33]     | -    | 76.6 96.3 99.2 | 61.4 88.9 95.1 | 56.6 84.5 92.0 | 39.2 68.0 81.3 |
| PFAN [41]    | -    | 76.5 96.3 99.0 | 61.6 89.6 95.2 | -    | -    | -    | -    |
| Unicoder-VL [19] | B | 84.3 97.3 99.3 | 69.7 93.5 97.2 | 62.3 87.1 92.8 | 46.7 76.0 85.3 |
| 12-in-1 [24] | B    | -    | -    | 65.2 91.0 96.2 | -    | -    | -    | -    |
| UNITER [5]   | B    | -    | -    | -    | -    | -    | -    | -    |
| UNITER [5]   | L    | -    | -    | -    | -    | -    | -    | -    |
| Oscar        | B    | 88.4 99.1 99.8 | 75.7 95.2 98.3 | 70.0 91.1 95.5 | 54.0 80.8 88.5 |
|              | L    | 89.8 98.8 99.7 | 78.2 95.8 98.3 | 73.5 92.2 96.0 | 57.5 82.8 89.8 |

(a) Image-text retrieval

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from [https://arxiv.org/abs/2004.06165](https://arxiv.org/abs/2004.06165)
## Experimental Results and Analysis

| Method   | VILBERT | VL-BERT | VisualBERT | LXMERT | 12-in-1 | UNITER$_B$ | UNITER$_L$ | OSCAR$_B$ | OSCAR$_L$ |
|----------|---------|---------|------------|--------|---------|------------|------------|-----------|-----------|
| Test-dev | 70.63   | 70.50   | 70.80      | 72.42  | 73.15   | 72.27      | **73.24**  | 73.16     | **73.61** |
| Test-std | 70.92   | 70.83   | 71.00      | 72.54  | –       | 72.46      | 73.40      | **73.44** | **73.82** |

### (b) VQA

| Method   | MAC | VisualBERT | LXMERT | 12-in-1 | UNITER$_B$ | UNITER$_L$ | OSCAR$_B$ | OSCAR$_L$ |
|----------|-----|------------|--------|---------|------------|------------|-----------|-----------|
| Dev      | 50.8| 67.40      | 74.90  | –       | 77.14      | **78.40**  | 78.07     | **79.12** |
| Test-P   | 51.4| 67.00      | 74.50  | 78.87   | 77.87      | **79.50**  | 78.36     | **80.37** |

### (c) NLVR2

| Method   | Test-dev | Test-std |
|----------|----------|----------|
| LXMERT [39] | 60.00    | 60.33    |
| MMN [4]    | –        | 60.83    |
| 12-in-1 [24] | –        | 60.65    |
| NSM [12]   | –        | 63.17    |
| OSCAR$_B$  | 61.19    | 61.23    |
| OSCAR$_L$  | **61.58**| **61.62**|

### (d) GQA

| Method   | cross-entropy optimization | CIDEr optimization |
|----------|-----------------------------|--------------------|
| BUTD [2] | 36.2 27.0 113.5 20.3       | 36.3 27.7 120.1 21.4|
| VLP [47] | 36.5 28.4 117.7 21.3       | 39.5 29.3 129.3 23.2|
| AoANet [11] | 37.2 28.4 119.8 21.3  | 38.9 29.2 129.8 22.4|
| OSCAR$_B$ | 36.5 **30.3** 123.7 23.1 | **40.5** 29.7 **137.6** 22.8|
| OSCAR$_L$ | **37.4** **30.7** 127.8 23.5 | **41.7** 30.6 **140.0** 24.5|

### (e) Image captioning on COCO

Xiuju Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from [https://arxiv.org/abs/2004.06165](https://arxiv.org/abs/2004.06165)
## Experimental Results and Analysis

| Method                        | in-domain | near-domain | out-of-domain | overall |
|-------------------------------|-----------|-------------|---------------|---------|
|                               | CIDEr     | SPICE       | CIDEr         | SPICE   |
| UpDown [1]                    | 78.1      | 11.6        | 57.7          | 10.3    | 31.3 | 8.3| 55.3 | 10.1|
| UpDown + CBS [1]              | 80.0      | 12.0        | 73.6          | 11.3    | 66.4 | 9.7| 73.1 | 11.1|
| UpDown + ELMo + CBS [1]       | 79.3      | **12.4**    | 73.8          | 11.4    | 71.7 | 9.9| 74.3 | 11.2|
| OscarB                        | 79.6      | 12.3        | 66.1          | 11.5    | 45.3 | 9.7| 63.8 | 11.2|
| OscarB + CBS                  | 80.0      | 12.1        | 80.4          | **12.2**| 75.3 | **10.6**| 79.3 | **11.9**|
| OscarB + SCST + CBS           | **83.4**  | 12.0        | **81.6**      | 12.0    | **77.6**| **10.6**| **81.1**| 11.7|
| OscarL                        | 79.9      | **12.4**    | 68.2          | 11.8    | 45.1 | 9.4| 65.2 | 11.4|
| OscarL + CBS                  | 78.8      | 12.2        | 78.9          | **12.1**| 77.4 | 10.5| 78.6 | **11.8**|
| OscarL + SCST + CBS           | **85.4**  | 11.9        | **84.0**      | 11.7    | **80.3**| 10.0| **83.4** | 11.4|

(f) Evaluation on NoCaps Val. Models are trained on COCO only without pre-training.

CBS- Constrained Beam Search  
SCST- Self-Critical Sequence Training

Xiujuan Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from [https://arxiv.org/abs/2004.06165](https://arxiv.org/abs/2004.06165)
Qualitative Studies

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
Qualitative Studies

**Oscar:** a small **train** on a city **street** with **people** near by.

**Baseline:** a **train** that is sitting on the side of the road.

**GT:** a small **train** on a city **street** with **people** near by.

A black and red small **train** in shopping area.

A group of **people** near a small railroad **train** in a mall.

**Tags:** sign, tree, sidewalk, **train**, woman, person, trees, **street**, bus, stairs, store, man, balcony, building, **people**

**Oscar:** a red **rose** and white **flowers** in a **vase**.

**Baseline:** a **vase** filled with red and white **flowers**.

**GT:** A red **rose** in a glass **vase** on a **table**

beautiful red **rose** and white **flowers** are in a **vase**.

The **bouquet** has one red **rose** in it.

**Tags:** leaf, **bouquet**, **flowers**, stem, **table**, **rose**, flower, leaves, **vase**, plant

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
Ablation Analysis

(a) VQA  
(b) Image Retrieval R@1  
(c) Image Captioning

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165, (2020). Retrieved from https://arxiv.org/abs/2004.06165
Key Strengths

- We can observe that the two different modalities are better aligned in the feature space visualization and OSCAR generates more detailed description of images than the baseline
Key Strengths

❖ We can observe that the two different modalities are better aligned in the feature space visualization and OSCAR generates more detailed description of images than the baseline
❖ OSCAR is highly parameter-efficient because the use of object tags as anchor points significantly eases the learning of semantic alignments between images and texts. OSCAR is pre-trained in 6.5 million pairs, which is less than 9.6 million pairs used for UNITER pre-training and 9.18 pairs for LXMERT
Key Strengths

❖ We can observe that the two different modalities are better aligned in the feature space visualization and OSCAR generates more detailed description of images than the baseline.

❖ OSCAR is highly parameter-efficient because the use of object tags as anchor points significantly eases the learning of semantic alignments between images and texts. OSCAR is pre-trained in 6.5 million pairs, which is less than 9.6 million pairs used for UNITER pre-training and 9.18 pairs for LXMERT.

❖ Techniques used in OSCAR for training and fine-tuning are similar to BERT. This makes it easier to come up with ideas to finetune for different V+L tasks. It is also easier to find documentation of BERT since it has good documentation on the internet.
Key Weaknesses

- In the Real-world images contain several novel objects unseen in training. Without ground-truth our model may not work well
Key Weaknesses

❖ In the Real-world images contain several novel objects unseen in training. Without ground-truth our model may not work well
❖ Collecting the image-caption training pairs can be very expensive process to train our model
Key Weaknesses

- In the Real-world images contain several novel objects unseen in training. Without ground-truth our model may not work well.
- Collecting the image-caption training pairs can be very expensive process to train our model.
- There is still a lot of overlap between different image regions passed to the model. We can add attention mechanism to the images passed to the model for better accuracy.
Future Work/ Open Research Questions

❖ Design a model that is able to caption novel/unseen objects while performing VL Tasks
❖ Train this model while attention on the image region so that we can further minimize ambiguity of the model.