# Evaluating Representation Learning with Reference Games

Anonymous ACL submission

## Abstract

TODO

## A Appendices

| Model | 100% | 50% | 25% | 10% | 5% | 1% |
|---|---|---|---|---|---|---|
| RNN | **86.08** | **83.60** | 81.72 | 77.57 | 75.35 | 60.59 |
| RNN+ | | | | | | |
| GloVe | 83.81 | 82.78 | 81.87 | **79.45** | **77.03** | 56.74 |
| Word2Vec | 83.81 | 82.13 | 81.78 | 79.43 | 71.23 | 56.33 |
| SkipThought | 82.39 | 79.15 | 79.00 | 75.44 | 69.76 | 56.75 |
| InferSent | 85.95 | 83.19 | **82.75** | 75.93 | 74.03 | 57.16 |
| BERT | 83.21 | 80.94 | 78.74 | 73.99 | 66.54 | 54.31 |
| GPT (OpenAI) | 82.86 | 79.62 | 75.93 | 69.33 | 66.24 | 48.74 |
| GPT2 | 84.22 | 80.85 | 79.30 | 75.44 | 63.22 | 53.64 |
| CTRL | 83.62 | 82.52 | 81.48 | 78.42 | 76.43 | **63.29** |
| Transformer-XL | 83.21 | 80.98 | 78.59 | 75.91 | 72.33 | 54.61 |
| XLNet | 80.98 | 79.24 | 78.07 | 73.80 | 67.04 | 57.08 |
| XLM | 80.01 | 76.54 | 72.83 | 70.86 | 60.20 | 50.72 |
| DistilBERT | 83.49 | 81.93 | 81.01 | 75.16 | 71.83 | 57.48 |
| RoBERTa | 83.68 | 80.03 | 79.04 | 69.74 | 63.87 | 52.45 |

Table 1: Evaluation of several language representations on the Colors in Context dataset. We vary the amount of training data used in transfer learning from 100% (30k examples) to 1% (300 examples).

| Model | 100% | 50% | 25% | 10% |
|---|---|---|---|---|
| RNN | | | | |
| GloVe | | | | |
| Word2Vec | | | | |
| SkipThought | | | | |
| InferSent | | | | |
| BERT | | | | |
| GPT (OpenAI) | | | | |
| GPT2 | | | | |
| CTRL | | | | |
| Transformer-XL | | | | |
| XLNet | | | | |
| XLM | | | | |
| DistilBERT | | | | |
| RoBERTa | | | | |

Table 2: Evaluation of several language representations on the ColorGrids in Context dataset. We vary the amount of training data used in transfer learning from 100% (2.3k examples) to 10% (230 examples).

| Model | Text | 100% | 50% | 25% | 10% | 5% | 1% |
|---|---|---|---|---|---|---|---|
| Vanilla | RNN | | | | | | |
| Vanilla | GloVe | | | | | | |
| Vanilla | Word2Vec | | | | | | |
| Vanilla | SkipThought | | | | | | |
| Vanilla | InferSent | | | | | | |
| Vanilla | BERT | | | | | | |
| Vanilla | GPT (OpenAI) | | | | | | |
| Vanilla | GPT2 | | | | | | |
| Vanilla | CTRL | | | | | | |
| Vanilla | Transformer-XL | | | | | | |
| Vanilla | XLNet | | | | | | |
| Vanilla | XLM | | | | | | |
| Vanilla | DistilBERT | | | | | | |
| Vanilla | RoBERTa | | | | | | |
| VGG19 | RNN | | | | | | |
| VGG19 | GloVe | | | | | | |
| VGG19 | Word2Vec | | | | | | |
| VGG19 | SkipThought | | | | | | |
| VGG19 | InferSent | | | | | | |
| VGG19 | BERT | | | | | | |
| VGG19 | GPT (OpenAI) | | | | | | |
| VGG19 | GPT2 | | | | | | |
| VGG19 | CTRL | | | | | | |
| VGG19 | Transformer-XL | | | | | | |
| VGG19 | XLNet | | | | | | |
| VGG19 | XLM | | | | | | |
| VGG19 | DistilBERT | | | | | | |
| VGG19 | RoBERTa | | | | | | |
| ResNet34 | RNN | | | | | | |
| ResNet34 | GloVe | | | | | | |
| ResNet34 | Word2Vec | | | | | | |
| ResNet34 | SkipThought | | | | | | |
| ResNet34 | InferSent | | | | | | |
| ResNet34 | BERT | | | | | | |
| ResNet34 | GPT (OpenAI) | | | | | | |
| ResNet34 | GPT2 | | | | | | |
| ResNet34 | CTRL | | | | | | |
| ResNet34 | Transformer-XL | | | | | | |
| ResNet34 | XLNet | | | | | | |
| ResNet34 | XLM | | | | | | |
| ResNet34 | DistilBERT | | | | | | |
| ResNet34 | RoBERTa | | | | | | |

Table 3: Evaluation (Part 1 of 2) of several multimodal representations on the Chairs in Context dataset.

| Model | Text | 100% | 50% | 25% | 10% | 5% | 1% |
|---|---|---|---|---|---|---|---|
| IR (ImageNet) | RNN | | | | | | |
| IR (ImageNet) | GloVe | | | | | | |
| IR (ImageNet) | Word2Vec | | | | | | |
| IR (ImageNet) | SkipThought | | | | | | |
| IR (ImageNet) | InferSent | | | | | | |
| IR (ImageNet) | BERT | | | | | | |
| IR (ImageNet) | GPT (OpenAI) | | | | | | |
| IR (ImageNet) | GPT2 | | | | | | |
| IR (ImageNet) | CTRL | | | | | | |
| IR (ImageNet) | Transformer-XL | | | | | | |
| IR (ImageNet) | XLNet | | | | | | |
| IR (ImageNet) | XLM | | | | | | |
| IR (ImageNet) | DistilBERT | | | | | | |
| IR (ImageNet) | RoBERTa | | | | | | |
| LA (ImageNet) | RNN | | | | | | |
| LA (ImageNet) | GloVe | | | | | | |
| LA (ImageNet) | Word2Vec | | | | | | |
| LA (ImageNet) | SkipThought | | | | | | |
| LA (ImageNet) | InferSent | | | | | | |
| LA (ImageNet) | BERT | | | | | | |
| LA (ImageNet) | GPT (OpenAI) | | | | | | |
| LA (ImageNet) | GPT2 | | | | | | |
| LA (ImageNet) | CTRL | | | | | | |
| LA (ImageNet) | Transformer-XL | | | | | | |
| LA (ImageNet) | XLNet | | | | | | |
| LA (ImageNet) | XLM | | | | | | |
| LA (ImageNet) | DistilBERT | | | | | | |
| LA (ImageNet) | RoBERTa | | | | | | |
| VAE (COCO) | RNN | | | | | | |
| VAE (COCO) | GloVe | | | | | | |
| VAE (COCO) | Word2Vec | | | | | | |
| VAE (COCO) | SkipThought | | | | | | |
| VAE (COCO) | InferSent | | | | | | |
| VAE (COCO) | BERT | | | | | | |
| VAE (COCO) | GPT (OpenAI) | | | | | | |
| VAE (COCO) | GPT2 | | | | | | |
| VAE (COCO) | CTRL | | | | | | |
| VAE (COCO) | Transformer-XL | | | | | | |
| VAE (COCO) | XLNet | | | | | | |
| VAE (COCO) | XLM | | | | | | |
| VAE (COCO) | DistilBERT | | | | | | |
| VAE (COCO) | RoBERTa | | | | | | |
| IR (COCO) | IR (COCO) | | | | | | |
| VAEVAE (COCO) | VAEVAE (COCO) | | | | | | |
| VAEGAN (COCO) | VAEGAN (COCO) | | | | | | |

Table 4: Evaluation (Part 2 of 2) of several multimodal representations on the Chairs in Context dataset.