

BigSleep を用いたテキストからの画像生成

情報メディア創成学類

芳賀 郁弥

2022/2/9

1 取り組んだタスク

BigSleep というネットワークモデルを用いてテキストから画像生成を行った。そして、最適化器の中で最も精度の高い画像を生成するものを調査した。

2 実験条件

- 実行環境
 - Google Colaboratory (無料)
 - python3.7
- データセット
 - CLIP (学習済み)
 - WIT (WebImageText) と呼ばれるデータセット
 - BigGAN (学習済み)
 - ImageNet
- エポック数: 10
- イテレーション数: 200
- シード: 0
- 学習率
 - adadelta: 50
 - SGD: 15
 - adam: 0.05
 - RMSprop: 0.05
 - adamW: 0.03
 - adagrad: 0.3

3 使用したネットワークの説明

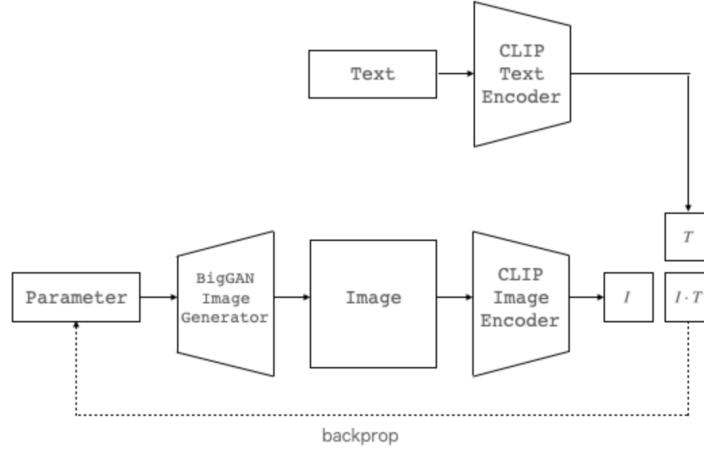


図 1: Big Sleep の構成.

実験を行うために, BigSleep [4] という学習済みネットワークを使用した. BigSleep は BigGAN [2] と CLIP [1] により構成されている. BigGAN は画像生成の役割を担う. また, CLIP は BigGAN から生成された画像及び入力テキストをそれぞれベクトル化して, 画像ベクトルとテキストベクトルのコサイン類似度が最大化するように誤差を伝える. [図 1]

4 実験結果

- 単語の単数形・複数形による生成される画像の違い (1週目)
- 最適化器を変えて画像生成 (2週目)
- コサイン類似度の遷移を可視化 (3週目)
- CLIP を用いた評価, DeepLabv3 を用いた評価 (4週目: 最終週)

以上の実験内容を本実験で取り組んだ. 2週目の実験に関して, デフォルトで最適化器は adam が使われている. そこで, adam に加えて adadelta, adagrad, adamW, RMSprop, SGD の合計 6 種類の最適化器を適用する. また, 「最適化器を変えて画像生成」に関しては他の実験でも行っているため, 4.1節では項立てて扱わない. さらに, 3週目の「コサイン類似度の遷移を可視化」は 4.2節の追加実験でも取り扱うため 4.1節では項立てない.

4.1 これまでの実験結果

4.1.1 単語の単数形・複数形による生成される画像の違い

「森の中にいるクマ」の画像を得るために, 入力テキスト「a bear in the forest」と「a bear in woods」から得られる画像の違いを比較した. この実験の目的は, wood (=木) の複数形 woods (=森) のように, 単数形と複数形の間にある意味の差異を表現できるかを検証することだ. 得られた画像は図 2 及び図 3 のようになった.

入力テキスト「a bear in woods」は, 主観的であるが, 入力テキストを表現できていない. 学習を進むごとに得られる画像の変遷からは「bear」と「woods」の関係性が意図したものになって



図 2: a bear in the forest

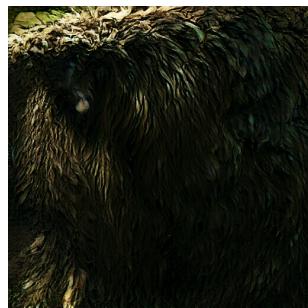


図 3: a bear in woods



図 4: a bear living in woods

いないように感じられた。

そこで、入力テキストを「a bear living in woods」と「living」を追記することで、「in」の意味の明確化を図った。結果、図4のような画像を得られた。図4は図3と比較して、「森の中にいるクマ」を画像化している。しかし、図2の「a bear in the forest」には及ばないように感じられる。

4.1.2 CLIP を用いた評価, DeepLabv3 を用いた評価



図 5: bicycle



図 6: chair



図 7: cat



図 8: motorbike

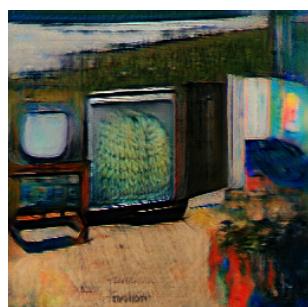


図 9: television

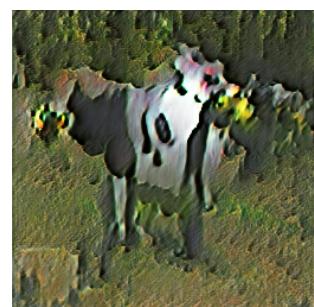


図 10: cow

入力テキスト chair, bicycle, cow, motorbike, television, cat の 6 種類から生成された画像 [図 5-10] を使用した。また、最適化器は adam を使用した。

CLIP を用いた評価

表 1: CLIP を用いた評価

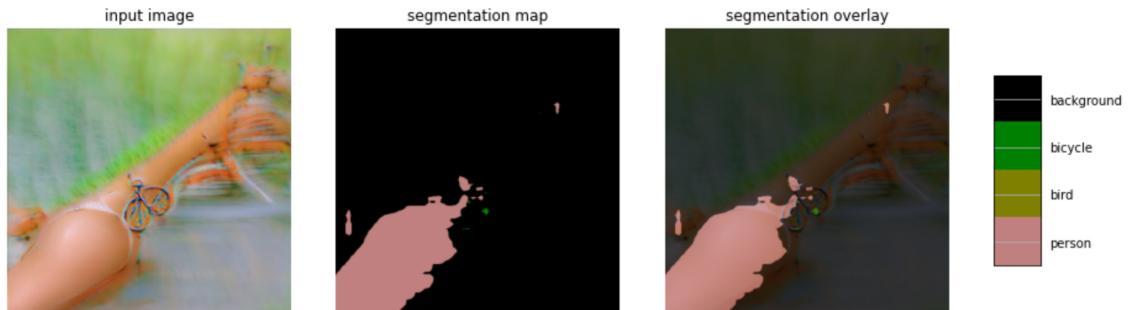
bicycle	chair	cat
bicycle: 99.76%	chair: 100.00%	tiger: 27.93%
motorcycle: 0.23%	table: 0.00%	girl: 12.01%
tractor: 0.00%	crab: 0.00%	mouse: 10.77%
lawn mower: 0.00%	can: 0.00%	boy: 7.28%
chair: 0.00%	apple: 0.00%	woman: 6.74%
motorbike	television	cow
motorcycle: 99.95%	television: 100.00%	cattle: 99.90%
bicycle: 0.05%	couch: 0.00%	bear: 0.01%
lawn mower: 0.00%	telephone: 0.00%	tractor: 0.01%
tractor: 0.00%	lamp: 0.00%	kangaroo: 0.01%
bowl: 0.00%	apple: 0.00%	lawn mower: 0.01%

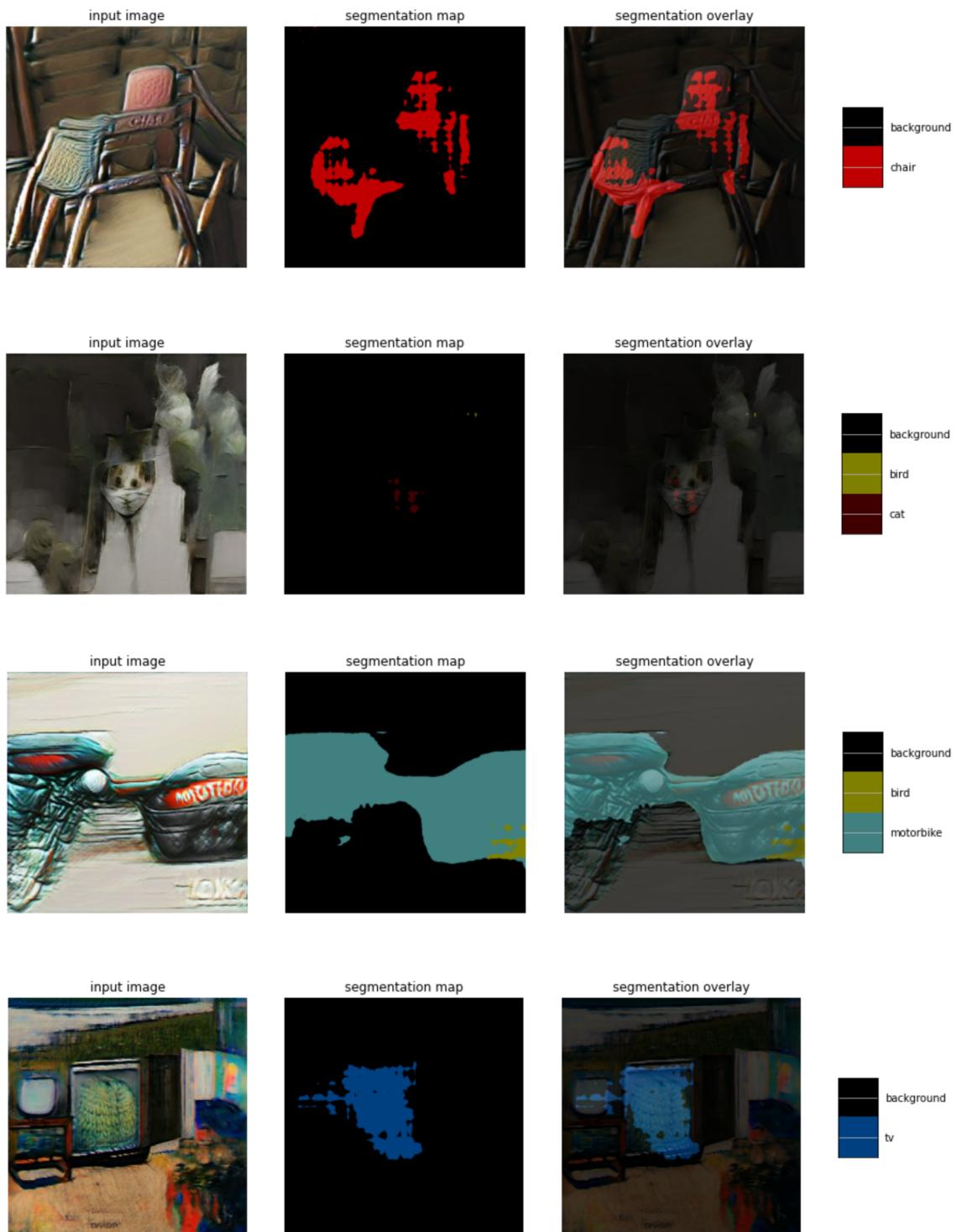
CLIPにおいて、テキストには CIFAR-100 のクラス名を用いた。CIFAR-100 はクラス数 100, 各クラス 600 枚の画像を持つデータセットである。この評価では、入力画像(図 5-図 10)と最も類似度の高い上位 5 種類のクラス名を出力した。また、ここで言う類似度とはテキスト特徴行列と画像特徴行列の積である。更に、softmax により値を加工した。結果は表 1 のようになった。

cat を除いた各生成画像は 100% に近い確率で生成するために入力したテキストと類似度が高いと CLIP により判定された。cat の精度が低い理由は、CIFAR-100 のクラスに cat というクラスがなかったためである。そこで、4.2 節の追加実験では CIFAR-100 に存在するクラス名で画像生成を行うことにした。

DeepLabv3 を用いた評価

DeepLabv3 [3] は入力画像の意味領域を判別する。図 11 から、bicycle, chair, cat, motorbike, television, cow によって生成された画像はそれぞれのテキストの特徴を持つことが確認された。





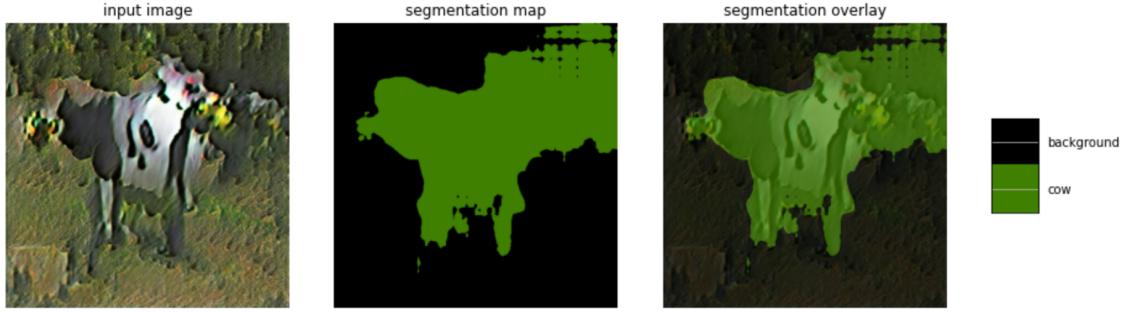


図 11: DeepLabv3 による bicycle, chair, cat, motorbike, television, cat の評価

4.2 追加実験

表 2: 6 種類の最適化器から生成された画像の CLIP を用いた評価

	adadelta	adagrad	adam	adamW	RMSprop	SGD
bicycle	35.56	39.09	40.59	38.19	39.56	37.78
chair	40.75	41.41	44.31	44.28	43.84	38.78
cattle	37.59	36.28	37.88	37.91	37.28	35.59
motorcycle	37.88	37.91	38.75	38.25	38.78	37.16
平均	37.95	38.67	40.38	39.66	39.87	37.33

入力テキスト bicycle, chair, cow, motorbike の 4 種類と最適化器 6 種類の組み合わせで、合計 24 枚の画像を Big Sleep により生成 [図 12-35] し、CLIP を用いてそれらの画像の定量評価を行なった。

結果は、全ての画像において画像生成の入力テキストと一致するテキストが CLIP では類似度が最も高いことが判明した。 (cow と cattle は牛という意味の単語、motorbike と motorcycle はバイクという意味の単語)

各画像の類似度 (テキスト特徴行列と画像特徴行列の積*100) を示したのが表 2 である。表 2 より、平均的に類似度の高い画像を生成した最適化器は順に adam, adamW, RMSprop, adagrad, adadelta, SGD であることがわかる。

入力テキスト: bicycle



図 12: adadelta



図 13: adagrad



図 14: adam



図 15: adamW



図 16: RMSprop



図 17: SGD

入力テキスト: chair

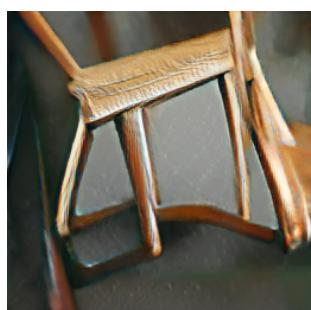


図 18: adadelta



図 19: adagrad

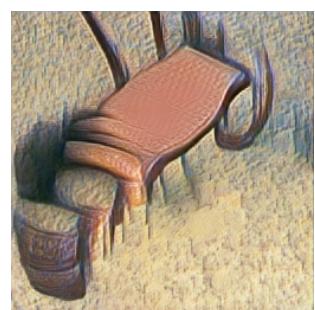


図 20: adam



図 21: adamW



図 22: RMSprop



図 23: SGD

入力テキスト: cow

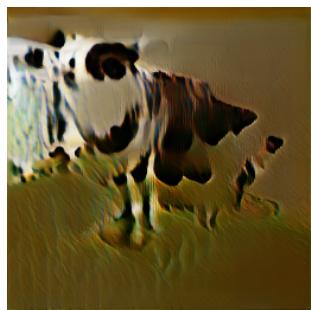


図 24: adadelta

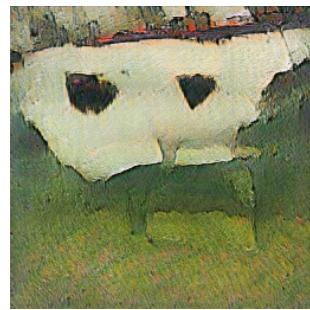


図 25: adagrad

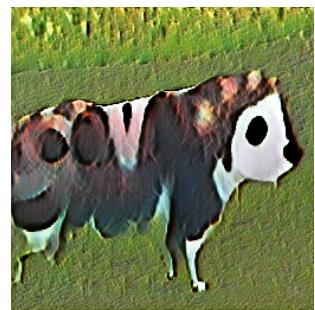


図 26: adam



図 27: adamW

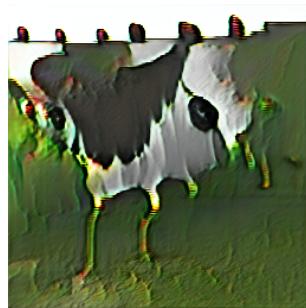


図 28: RMSprop

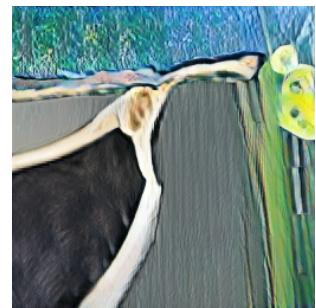


図 29: SGD

入力テキスト: motorbike



図 30: adadelta



図 31: adagrad



図 32: adam



図 33: adamW



図 34: RMSprop



図 35: SGD

5 追加結果の考察

追加実験で行った実験について考察を行う。図 36 は追加実験において、入力テキスト motorbike で画像を生成したときの損失の遷移である。図 36 より、損失の大小が画像の精度に影響を与えているとは言い難いと言える。なぜならば、表 2 より、最適化器 adam で入力テキスト motorbike により生成された画像は類似度が 38.75 であり、最適化器 adamW で入力テキスト motorbike により生成された画像は類似度が 38.25 という結果が出た。しかし、図 36 の損失の遷移では最適化器 adam が -0.36 付近で収束しているのに対して、最適化器 adamW は -0.42 付近で収束している。よって、損失の収束値が画像の精度に影響を与えているとは言えない。

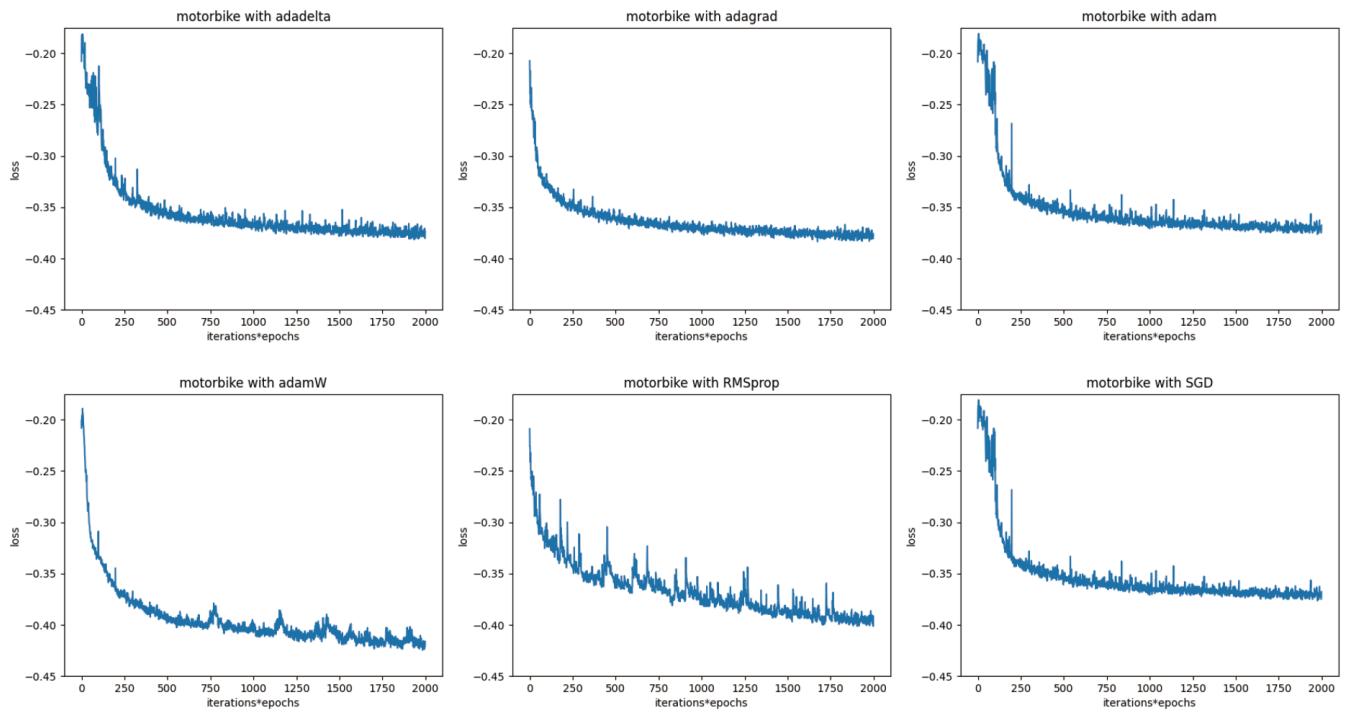


図 36: 各最適化器における入力テキスト motorbike の損失の遷移

6 使用したソースファイル

Big Sleep 実行場所 (google colab)

https://colab.research.google.com/drive/13-IBJCKDmiwdamCRx93EhHpIQBMuweM4#scrollTo=0GP8cMdoP_hw

Big Sleep ソースファイル

https://drive.google.com/drive/folders/1BwPxmqpK2VM4F7v2d146l8HCLaL1T8_-?usp=sharing

DeepLabv3 実行場所 (google colab)

<https://colab.research.google.com/drive/1wlP1srXHTf3fSSe3X8YmfAmwBSjAFXMB?usp=sharing>

CLIP 実行場所 (google colab)

<https://colab.research.google.com/drive/1URLQop0zeV4HKp6RTGzLu60t58N3Eot0?usp=sharing>

参考文献

- [1] Alec Radford, et al., Learning Transferable Visual Models From Natural Language Supervision, 2021. <https://arxiv.org/pdf/2103.00020.pdf>
- [2] Andrew Brock, et al., LARGE SCALE GAN TRAINING FOR HIGH FIDELITY NATURAL IMAGE SYNTHESIS, 2019. <https://arxiv.org/pdf/1809.11096v2.pdf>
- [3] Liang-Chieh Chen, et al., Rethinking Atrous Convolution for Semantic Image Segmentation, 2017. <https://arxiv.org/pdf/1706.05587.pdf>

[4] <https://github.com/lucidrains/big-sleep>