Generating cartoons from text using NLP and image processing

Aubhishek Zaman, PhD


CREATED USING
POWTOON

Strategies

1. Use a natural language processing (NLP) parts of speech algorithm for user input to call images using text-based image labels (located in file name):
2. Use a pretrained model (e.g., CLIP, GLIDE, BIOBERT and DALL-E) for both text to image generation
3. De-novo training of an image and text embedding followed by fusing them onto an adversarial network architecture

Methodologies

1) Collect and preprocess data
2) Train an NLP model
3) Train a computer vision model
4) Design an architecture that combines the textual and visual information.
5) Evaluate and refine

Labeled Image with image name matching algorithm with noun and verb isolation was functional

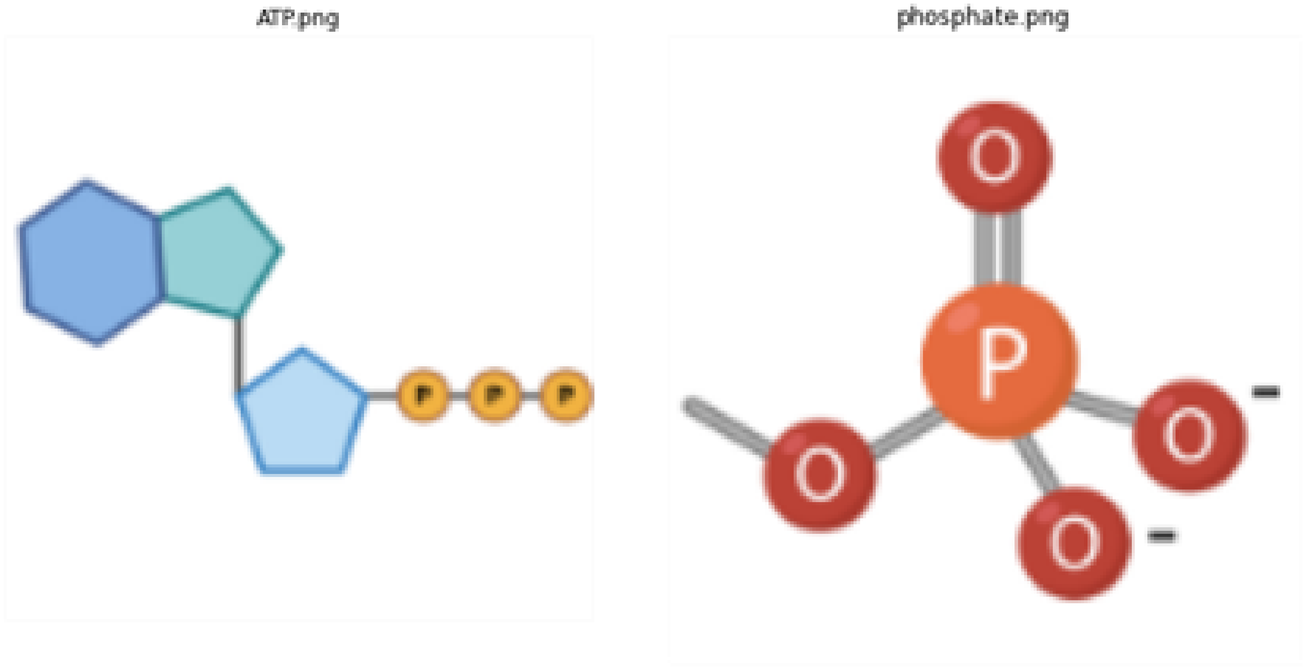


Figure 1: Image generation on canvas for text input - 'ATP and phosphate'

Required granularity is inconsistent with pre-tarained models

a

	a girl wearing a beanie	a chemical structure
image1	0.00	100.00
image2	0.00	100.00

b

	ATP	Phosphate
image1	22.30	77.70
image2	1.79	98.21

Figure 2: Granularity of image classification. (a) Image granularity was better for subjects with distinct features but fails (b) for images with chemical structure variations

De-novo image classification is 70% accurate without required granularity

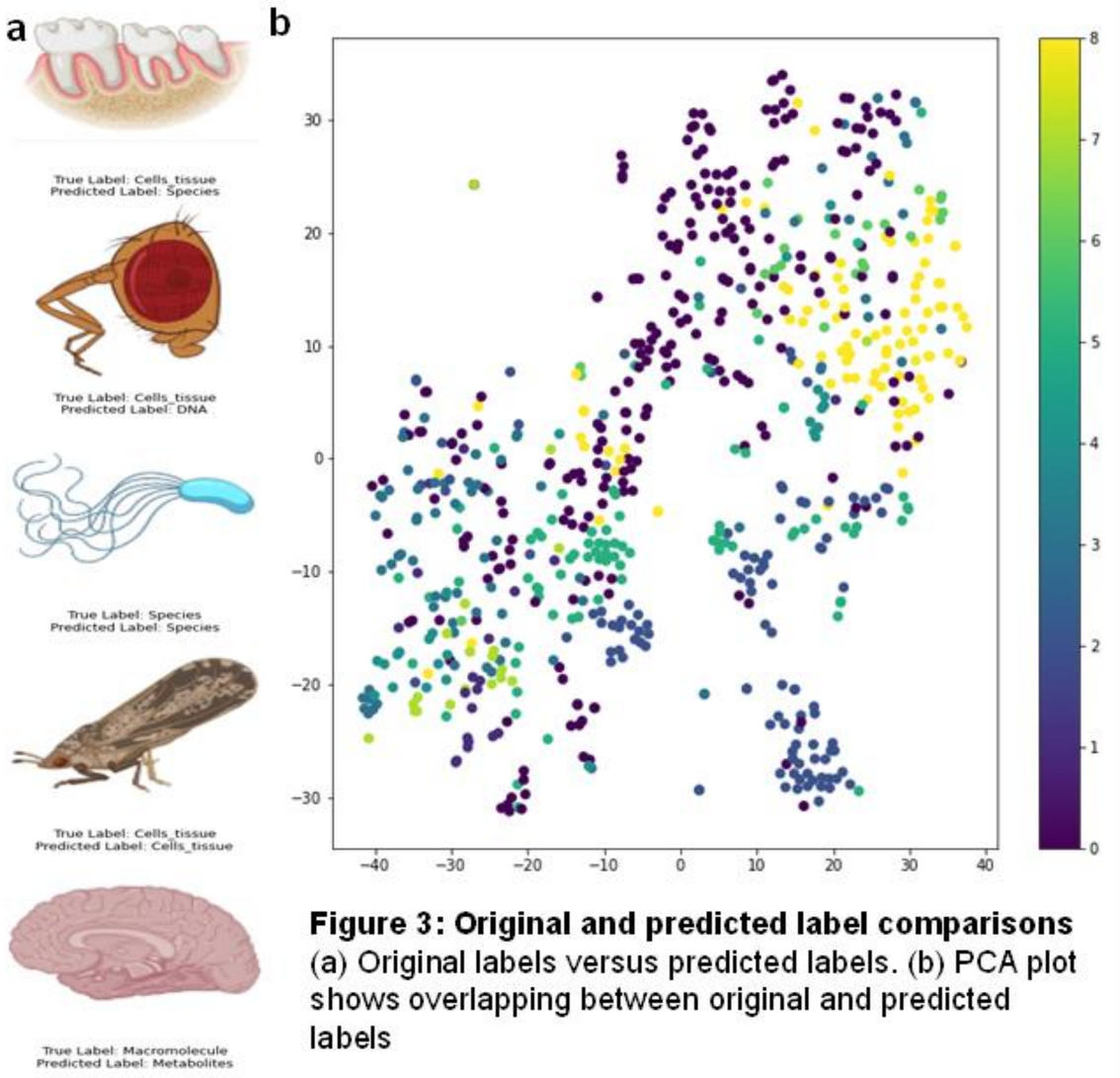


Figure 3: Original and predicted label comparisons (a) Original labels versus predicted labels. (b) PCA plot shows overlapping between original and predicted labels

De-novo text classification is predominant with prepositions and need to be

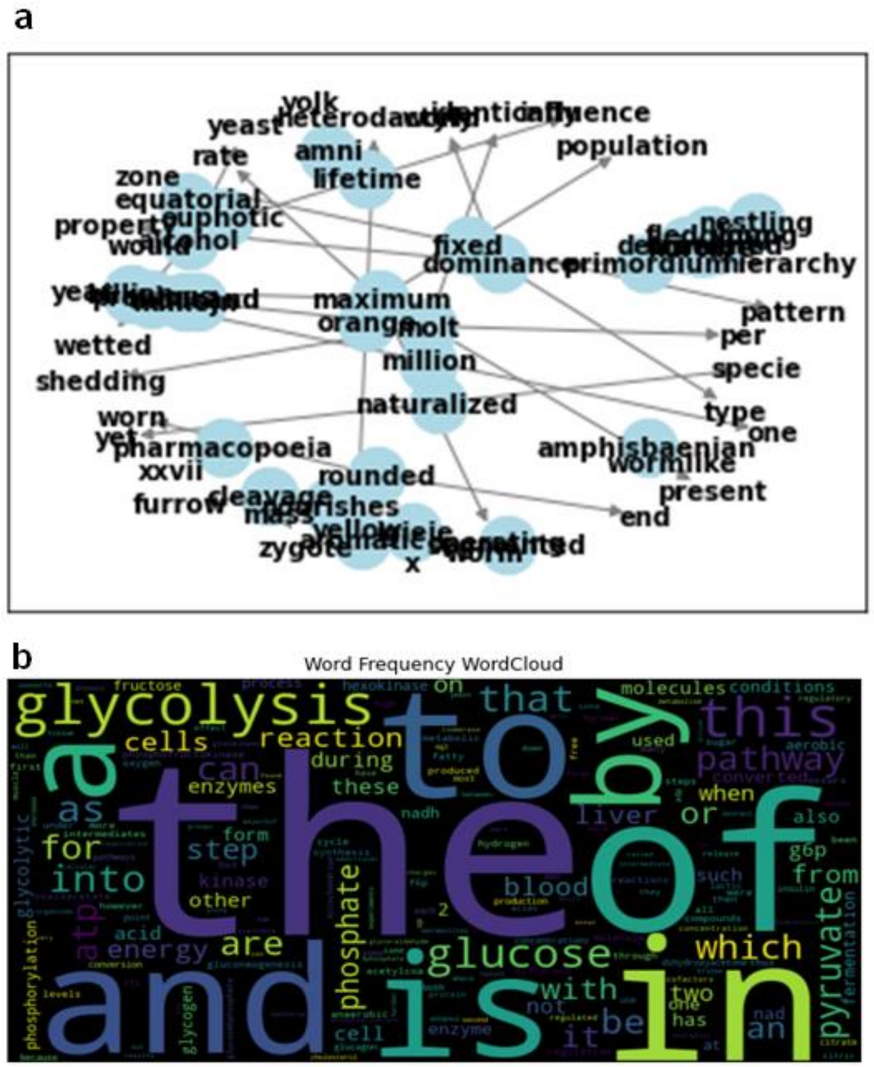
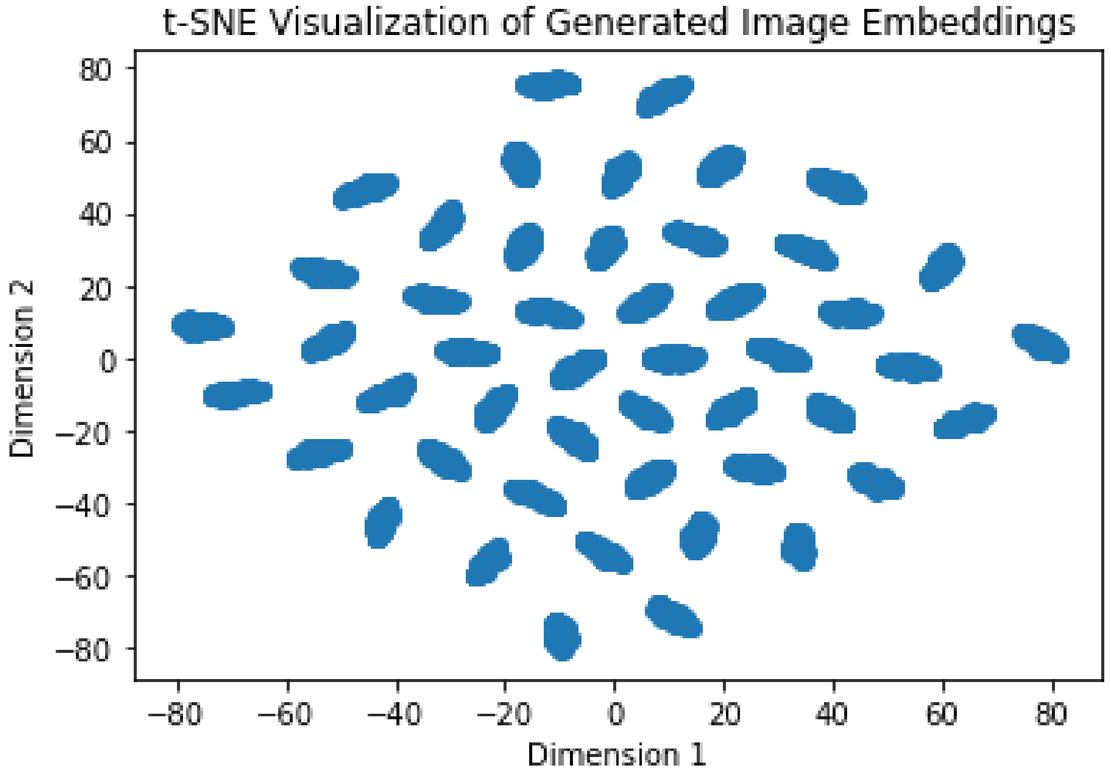


Figure 4: Texts embeddings using dictionary (a) and beautiful soup (b).

Fused embeddings fail to generate any images with generative adversarial network



A priority and exceptions -based NLP algorithm that matches labeled image name with text input

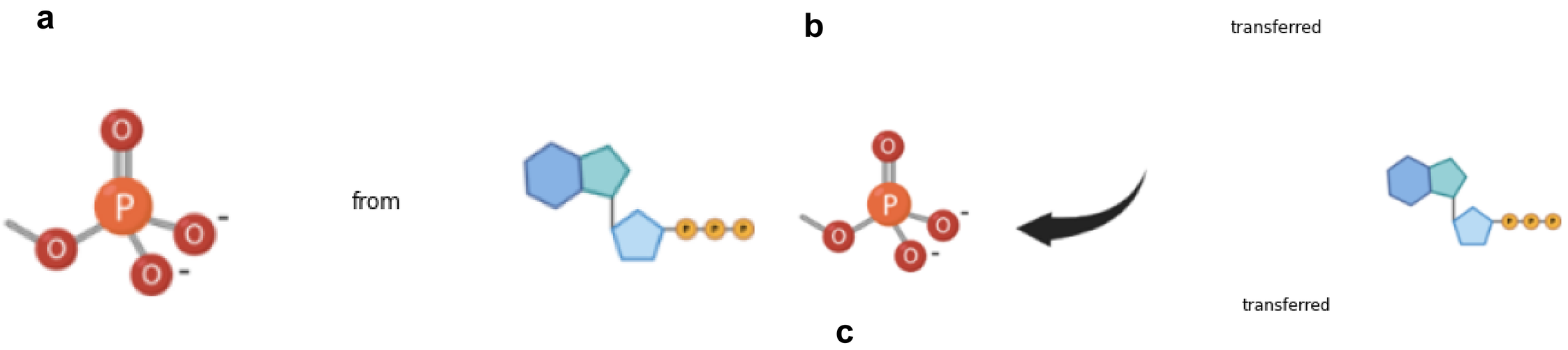


Figure 4: Generated cartoons-evolution of algorithm. The text input phosphate was generated from ATP (a) without any verbs as a clipart (b) with verbs where from and to and its synonyms were artificially labeled with arrow. (c) Multiple images were recommended for user input, (b) was the first and (c) was the second recommendation of lesser quality

Additional use cases

Figure 6: Emojis and clipart for input text (sad)

Future directions

- Further databases.
- Bigger datasets; for more granular task and hence need to be further tuned. For each imaged non-overlapping labeling labeling would be necessary.
- Also, multiple images for the same label is needed.
- For text embeddings, limiting training words to nouns and verbs