Modelling andGeneratingFakeWebsites for CyberDeception

Michael Longland
Project Goal

- Analyse the capability of deep learning models to generate reasonable looking websites
- Generate realistically structured documents
- Generate realistic between-page link networks
- Combine the generation of documents and link structure to make realistic fake websites
Motivation

• Text generation – highly active research area
  – Extending to structured documents
  – Extending to multiple documents

• Application: Network defence
  – Honeypots
Approach

• Data collection – Wikipedia articles
  – Well-structured documents, not overly-complex structure
  – Can use to later extend to documents with more complex structure
    • HTML documents
• Article generation – GPT-2
  – GPT-2 ‘Generative Pretrained Transformer 2’
• Link network generation – GRAN
  – GRAN, ‘Graph Recurrent Attention Networks’
• Linking GPT-2 and GRAN to make fake mini-Wikipedias
Related Work

- Transformer language models
  - BERT, XLNet, Transformer-XL, etc.
  - Transformer-XL trained on Wikipedia text but doesn’t look at generation of entirely new articles
  - None of these consider generation of multiple linked samples/documents

- Cyber deception, network defence – honeypots, honeyfiles
Data Collection

- Data collection from Wikipedia:
  - ~30,000 articles from the categories, subcategories and sub-subcategories of Military of Australia, Military of New Zealand, Military of Canada, Military of the United Kingdom, Military of the United States
    - Page content
    - List of links to other Wikipedia pages
    - Selection of categories was somewhat arbitrary
- Data cleaning:
  - Removed articles with > 1023 words so model learns articles should be this length or under
    - GPT-2 samples (generated articles) are limited to 1024 words
  - ~21,000 articles remain after this, 49MB raw text
Text Generation – GPT-2

- GPT-2 is a transformer model for language generation
- Pretrained GPT-2 models finetuned on the Wikipedia article dataset
- Articles in training data prepended with '== Article Start==‘
- Samples generated using only '== Article Start==‘ as prompt
  - Model learns to generate a title then the rest of the article from only this
- Compared two pretrained model sizes, 124M and 345M parameters
  - 124M: faster to train and sample from, generated articles less complex in vocabulary and structure, more likely to have contradictions
  - 345M: slower to train and sample from, significantly better in all other aspects
Text Generation – GPT-2

- Finetuned model usually generates coherent articles
- Generated articles sections are ordered correctly – a title and intro section, then body sections/subsections, then a reference list and/or external links
- Generated article structure is similar to structure in real articles
- Section structure: sections and subsections relate to the article title and follow logically
  - Wikipedia uses a lightweight markup for article structure, e.g.
    - == denotes a section heading
    - === denotes a subsection heading
    - ==== denotes a sub-subsection heading
  - Model learns these
Example generated article:

== Article Start ==
North American Mobile Force


== History ==
The Mobile Security Force consisted of approximately 40 soldiers, two trucks and two smaller vehicles and was commanded by Colonel Robert H. Clausen. The unit was initially trained in the same manner as the Army Mobile Security Force, namely, with the Mobile Security Force Command and Headquarters, coming into existence on 30 July 1969. The unit was assigned to the 4th Infantry Division, and subsequently to the 1st Light Anti-Tank Regiment and the 1st Battalion, 130th Armor Regiment.

== Operations ==
The unit was intended to deploy for operations in Vietnam, but was not sent. The unit was commanded by Colonel Clausen, who had the rank of Colonel. The unit consisted of four companies (each consisting of two trucks and two smaller vehicles) and one company of reserve soldiers. Clausen provided command and control of the Mobile Security Force. The unit was tasked with the mission of providing rapid response to the inter-war period and to provide force protection to the division. The unit remained on active duty until 20 March 1971, during the Vietnam War, when the unit was returned to US Army Reserve. In June 1971, and the first time the unit had been activated, the Mobile Security Force was deployed to the Philippines and Romania.

== See also ==
United States Army in Southeast Asia
North American advisers’ unit
<|endoftext|>
Text Generation – GPT-2

• Paragraphs are relevant to the section heading, stay on topic throughout, and usually correctly reference statements from earlier paragraphs

== Article Start ==

Albert Le Roux

Albert Le Roux (August 14, 1879 – July 16, 1961) was a Canadian historian and political activist. He served in the Canadian Infantry during the First World War.

== Early life and education ==

Le Roux was born in 1879 in Ridgeway, Ontario, the son of Alphonse Le Roux. As a young man, Le Roux attended Mount Allison School in Toronto until his father left the family when Le Roux was about twelve. He was subsequently educated at Osgoode Hall and Mount Allison College.

== Military career ==

Le Roux entered the Canadian Infantry in the Second Canadian Expeditionary Force in 1916. Le Roux was commissioned in the men's 1st Battalion in January 1917, and was seconded to the 2nd Battalion. Le Roux was promoted to lieutenant in November 1917 and captain in November 1918. In 1917, Le Roux was in command of a battalion in the Italian Campaign. Le Roux was promoted to captain in May 1919, and was present at the Battle of the Somme when Thomas Hopkinson's Battalion was sent to the front as reinforcements. Le Roux was also present at the Battle of the Castine, Battle of Saint-André-sur-Richelieu, Battle of the Somme, Battle of the Loire and Battle of Ligny-sur-Seull. He was next awarded the Croix de Guerre with Bars. He was also promoted to brigadier-general in March 1921. Le Roux was promoted to colonel in August 1921. Le Roux served as adjutant general of the Canadian Corps in the First World War. He was also promoted to major general in November 1922.

== Personal life ==

Le Roux lived in Toronto, Ontario, until his death in 1961 at age 61. He was married to Aline Toplinden.

<|endoftext|>
Text Generation – GPT-2

- Realistic book titles and URLs in the references and external links sections, and are generated only in these sections
- Books, URLs very rarely actually exist

http://www.af.mil/af/aviation_service/service/wps_sas/
http://www.nps.gov/landbirds/Wyoming/most/80329b08b.html
http://www.fordeasternuh.org/british-scots-catalogue-experiments/B1/Map_0/eu1.pdf

Link Network Generation – GRAN

- Training data: subgraphs of the Wikipedia link network.
- Sampled order 1 egocentric networks with between 15-100 nodes:
  1. A node (article) is selected
  2. The articles it links to are selected, provided these articles exist in the training data, with edges denoting a link between them
  3. The edges (links) between nodes added in step 2 are included
- Node degree distribution approximately follows a power law, the max degree is ~700, and the most have degree <100. Also not interested in larger networks.
- GRAN can’t include node attributes (e.g. article title), these were removed
Link Network Generation – GRAN

• Also performs well
• Captures network attributes seen in training graphs:
  – Samples are always egocentric
  – High clustering/large cliques in some samples
  – Network size, degree distribution similar
• Samples aren’t identical to training graphs
Training data

GRAN samples
Mini-Wikipedia Generation

Method:

• Generating a network with $N$ nodes, generating $N$ articles, then mapping these articles to nodes to maximise the total cosine similarity across edges

• Smoothed inverse frequency (SIF) embedding applied to articles
  – Paragraphs represented by 300-dimensional vector
  – Common words contribute less to the vector

• Total similarity $S = \text{sum of the cosine similarity between linked nodes}$

• Optimal assignment of articles to nodes can be written as a case of the quadratic assignment problem – known to be NP-Hard

• Heuristic: greedy similarity maximisation
Mini-Wikipedia Generation

Greedy similarity maximisation:

1) Randomly assign articles to nodes, calculate total similarity $S$
2) For each pair of nodes, calculate the new $S$ if they were swapped
3) Keep the swap which increases $S$ the most
4) Repeat (2)-(3) until $S$ can no longer increase
5) Repeat (1)-(4) ten (or more) times, select the best run
Future Work

- Generating HTML documents
- Including relevant images or other media in the generated documents
- Using newer transformer models – many more have been published and released since GPT-2
- Different methods of selecting subgraphs for link networks
- Different similarity measures or algorithms for assigning articles to nodes
- Simultaneously generating articles and the link network
References

Questions?