**Matriculation Number: 00503319, 00616219**

**Literature Survey: Visual-to-audio aid for visually impaired**

**Summary:**

In our literature survey for this research, we did not come across the exact project idea as a computational technology perspective but ideas surrounding the topic and specific modules of it. We focused on the major components of our research project which are image captioning and sematic text comparison. Both the modules have been thoroughly researched by various authors which helped us analyze the pros and cons of different methodologies. Some of the writings proposed enhancements that can be used to provide better results as well.

Considering all the papers we found relating to image captioning and semantic text similarity were deeply studied for any insights to be gained, and finally we selected 6 of those papers which were most insightful with regards to our idea. We also came across the key methodologies and players in the field for various components and the idea as-a-whole which will be described here.

**Search Engines:** IEEE, DBLP, ELSEVIER, Google Scholar, arXiv.

**Field Scope:**

The field of providing aid to visually impaired with the means of other human senses was first presented through the idea called *Sensory Substitution* coined by *Paul Bach-y-Rita et al*. The idea was initially generic with regards to choosing the sensory modality for compensating the loss of another. Later on *Paul Bach-y-Rita* came up with a tactile approach, called the *Tactile-visual-Sensory-Substitution* to convert images to tactile sensations on sensors on the body. Later a more popular approach with sound was introduced by Peter Meijer with his invention of *The vOICe.*

In the future the sensory substitution field will move towards sensory augmentation and brain implants to compensate the loss of a sensory modality.

**Key Players:**

There are some key players who have researched with the same motivation as ours to help the visually impaired but have dominantly taken a completely different approach involving bio-mechanics to go for sensory substitution devices (SSD). The most notable of them would be The vOICe SSD, Tactile Vision Sensory Substitution (TVSS), Prosthesis Substituting Vision for Audition (PSVA).

**Key Methodologies Identified:**

|  |  |
| --- | --- |
| **IMAGE ENCODERS** | AlexNet, InceptionV3, GoogleNet, VGGNet, ResNet |
| **LANGUAGE DECODERS** | LSTM, GRU, Language CNN |
| **DATASETS** | flickr 8k/30k, MS-COCO, imagenet |
| **EVALUATION METRICS** | BLEU, METEOR, CIDEr |

**Literature Review:**

|  |  |  |  |
| --- | --- | --- | --- |
| Sr no | Research Paper | Author | Inference |
| 1 | Compositional generalization in a deep sequence to sequence model | Jake Russi, Jason Jo, Randall O’Reilly, Yoshua Bengio | Significant improvement on generalization task by imposing an analogous separation on SCAN dataset as compared to previous works |
| 2 | Image Captioning with Semantic Attention | Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo | Top-down and bottom-up approach is used which is than coupled with RNN to attend on rich semantic attributes. The proposed method gives state-of-art performance across popular benchmark |
| 3 | Short Text Similarity with Word Embeddings | Tom Kenter, Maarten de Rijke | Computing semantic similarity between short text using semantic only method outperforms previous approaches. The proposed method is independent of NLP tools and hence can be applied to varied domains and languages |
| 4 | Analysis of diversity-accuracy trade off in image captioning | Ruotian Luo, Gregory Shakhnarovich - TTI-Chicago | Diverse beam search has best trade off but slowest, naive sampling with low temperature is competitive. The existing metrics do nit measure the trade off rather diversity or accuracy distinctly |
| 5 | Show, Attend and Tell: Neural Image Caption Generation with Visual Attention | Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio | Attention model significantly improves caption accuracy. Soft attention (deterministic) is more popular as back propagation is more effective on regular and easier gradient whereas hard attention (deterministic) depends on number if sampling and how well it is sampled |
| 6 | LSTM vs. GRU vs. Bidirectional RNN for script generation | Sanidhya Mangal, Poorva Joshi, Rahul Modak | LSTM are most efficient having the capability to forget irrelevant data, GRU is a bit less accurate however is faster, bidirectional RNN are the most accurate however are the slowest |