We build a joint multimodal model of text and images for automatically assigning illustrative images to journalistic articles. We approach the task as an unsupervised *representation learning* problem of finding a common representation that abstracts from individual modalities, inspired by multimodal Deep Boltzmann Machine of Srivastava and Salakhutdinov. We use state-of-the-art image content classification features obtained from the Convolutional Neural Network of Krizhevsky et al. as input “images” and entire documents instead of keywords as input texts. A deep learning and experiment management library *Safire* has been developed. We have not been able to create a successful retrieval system because of difficulties with training neural networks on the very sparse word observation. However, we have gained substantial understanding of the nature of these difficulties and thus are confident that we will be able to improve in future work.