Vytváříme společný pravděpodobnostní model textu a obrázků pro úlohu automatického přiřazování ilustračních fotografií k novinovým článkům. Přistupujeme k úloze z hlediska *učení reprezentací*: chceme nalézt společnou reprezentaci textu i obrázků nezávislou na vlastnostech jednotlivých modalit, podobně jako multimodální hluboký Boltzmannův stroj Srivastavy a Salakhutdinova. Vstupní obrázky reprezentujeme pomocí předposlední vrstvy konvoluční neuronové sítě Krizhevského a kol., state-of-the-art reprezentace obrázků na základě jejich obsahu. Vytvořili jsme knihovnu *Safire* pro hluboké učení a správu multimodálních experimentů. Úspěšný vyhledávací systém se nám vyvinout nepodařilo, kvůli obtížnému trénování neuronových sítí na velmi řídkých textových datech. Porozuměli jsme však povaze těchto potíží tak, že věříme, že v navazující práci můžeme lepších výsledků dosáhnout.